Confidence Intervals for Expected Abundance of Rare Species

David FLETCHER and Malcolm FADDY

In many ecological research studies, abundance data are skewed and contain more zeros than might be expected. Often, the aim is to model abundance in terms of covariates, and to estimate expected abundance for a given set of covariate values. An approach that has been advocated recently involves the use of a conditional model. This allows one to separately model *presence* and *abundance given presence*, which should lead to a more complete understanding as to how the covariates influence abundance. The focus of this article is on the calculation of confidence intervals for expected abundance given particular values of the covariates. The standard Wald confidence interval is symmetric, and therefore unlikely to be of much use for skewed data, where reliable confidence intervals for abundance will generally be asymmetric. The purpose of this article is to show how to calculate a profile likelihood confidence interval for expected abundance using a conditional model.

Key Words: Conditional model; Negative binomial; Profile likelihood; Skewness; Zero-inflation.

1. INTRODUCTION

In many ecological research studies, abundance data often exhibit two features: a substantial proportion of the values are zero, and the remainder have a skewed distribution. In many types of study, the aim is to model the abundances in terms of one or more covariates, and to estimate the expected abundance for a given set of covariate values. We consider the use of a conditional-model approach for this purpose, as advocated by Welsh, Cunningham, Donnelly, and Lindenmayer (1996). This involves separately modeling (a) the presence of the species, and (b) the abundance of the species given that it is present (hereafter called the conditional abundance). This approach has the advantage that we can model these two aspects of the data separately, and thereby gain insight into whether they are being influenced by the covariates in different ways. Welsh et al. (1996) suggested using a truncated Poisson or truncated negative binomial distribution for the positive abundances. Dobbie and Welsh (2001) extended this type of model to deal with serial dependence in

David Fletcher is Associate Professor, Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin, New Zealand (E-mail *dfletcher@maths.otago.ac.nz*). Malcolm Faddy is Adjunct Professor, School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland 4001, Australia.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 12, Number 3, Pages 315–324
DOI: 10.1198/108571107X229322

repeated measurements, while Barry and Welsh (2002) advocated the use of conditional generalized additive models. Martin et al. (2005) provided a useful review of the issue of zero-adjustment in ecological data.

Estimation of the expected abundance for a given set of covariate values is straightforward using a conditional model, and is based on the maximum likelihood estimates of the parameters from the two-component models. Welsh et al. (1996) suggested use of a Wald confidence interval around the estimate of the expected abundance. This interval is symmetric, being based on the assumption that the estimate has a sampling distribution that is approximately normal. The skewness we usually observe in abundance data suggests that an interval which contains the true expected abundance 95% of the time is likely to be asymmetric relative to the estimated expected abundance. Profile likelihood confidence intervals (Venzon and Moolgavkar 1988) are generally not symmetric and are therefore more informative than Wald intervals. The asymmetry will reflect differential uncertainty in the estimation, by having a different range of values on one side of the maximum likelihood estimate compared with the other. Such information is potentially useful, and is clearly not available from a symmetric confidence interval, such as the Wald interval. In addition, the Wald interval can be viewed as an approximation to the profile likelihood confidence interval (Brown, Cai, and DasGupta 2003; Cox and Hinkley 1974, p. 343; Lambert 1992). This was all recognized by Welsh et al. (1996), but they claimed that "... it is difficult to see how to apply this method [*profile likelihood*] to calculate confidence intervals for predicted values from the integrated model." The purpose of this article is to show how to calculate such intervals.

2. MOTIVATING EXAMPLE

The motivation for this article came from the first author's analysis of data from a study carried out by Eduardo Viloutta of the Department of Conservation, New Zealand. He was interested in assessing the relationship between abundance of a seaweed (*Ecklonia radiata*; hereafter *Ecklonia*) and that of a sea urchin (*Evechinus chloroticus*; hereafter *Evechinus*) in Fiordland, New Zealand. This particular species of sea urchin is found only in New Zealand and on subantarctic islands. The motivation for the study came from assessing the potential impact on the abundance of *Ecklonia* of establishing an *Evechinus* fishery. We therefore considered abundance of *Ecklonia* to be the response variable, with abundance of *Evechinus* being a predictor variable. The data we use to illustrate the method are part of a larger dataset, including a number of covariates that were also thought to influence the abundance of *Ecklonia*. We do not consider these here, for ease of presentation: the profile likelihood method can be applied quite generally, regardless of the number of predictor variables. Further information on the study and the complete dataset can be found in Fletcher, MacKenzie, and Villouta (2005). The data we use here are shown in Table 1.

At each of 103 locations, abundance of both *Ecklonia* and *Evechinus* was measured using a 25m^2 quadrat. There were 27 locations that showed 0 counts of *Ecklonia*, and Figure 1 shows a scatterplot of the abundances of the two species (the circled observation is discussed in Section 4).

Table 1. Data on the relationship between abundance of *Ecklonia radiata* (plants/quadrat) and *Evechinus chloroticus* (individuals/quadrat), measured using a $25m^2$ quadrat at each of 103 locations.

Ecklonia	*Evechinus*	*Ecklonia*	*Evechinus*	*Ecklonia*	*Evechinus*
0	3	2	0	46	28
0	29	3	14	48	4
0	38	3	115	49	0
0	0	6	22	56	29
0	18	8	53	57	2
0	6	8	15	57	15
0	18	11	64	57	11
0	11	12	28	58	0
0	19	12	52	58	28
0	32	15	14	59	0
0	33	16	0	59	9
0	133	19	11	61	0
0	19	20	11	64	1
0	0	20	20	67	0
0	13	20	14	70	0
0	5	22	7	72	0
0	5	24	5	73	5
0	27	25	0	79	2
0	39	26	12	81	3
0	110	28	20	81	8
0	31	30	11	84	0
0	50	30	11	87	0
0	0	31	20	88	25
0	34	33	13	89	2
0	220	33	52	94	28
0	78	33	19	96	2
0	27	33	29	96	4
1	23	34	24	103	29
1	15	35	25	118	13
1	15	38	0	137	0
2	35	38	19	157	4
2	34	39	0	168	5
2	15	40	35	203	50
2	19	44	0		
2	16	45	0		

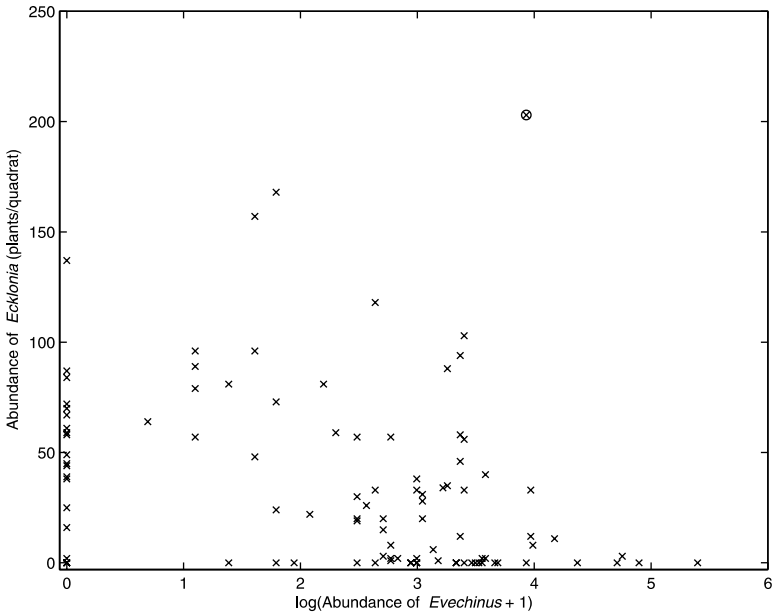

Figure 1. Scatterplot of the abundance of *Ecklonia* and the abundance of *Evechinus*.

The objective of the analysis is to estimate the expected abundance of *Ecklonia* (plants/quadrat) for a given abundance of *Evechinus* (individuals/quadrat). Preliminary analyses suggested an approximately linear relationship between *Ecklonia* and log(*Evechinus* +1), so we used the latter as our predictor variable in each of the two-component models.

3. PROFILE LIKELIHOOD INTERVAL FOR THE CONDITIONAL MODEL

Let Y be the abundance of the species of interest (response variable) and let Z indicate presence (1) or absence (0) of the species. We model Z and $Y|Z = 1$ in terms of covariates $\mathbf{x} = (x_0, x_1, \ldots, x_p)^{\mathrm{T}}$ and $\mathbf{w} = (w_0, w_1, \ldots, w_q)^{\mathrm{T}}$, respectively. Note that we will usually wish to include an intercept in each of these models, and so we have $x_0 = w_0 = 1$. If we assume a truncated negative binomial distribution for $Y|Z = 1$, we have

$$\Pr(Y = 0|\mathbf{x}) = \Pr(Z = 0|\mathbf{x}) = 1 - \pi(\mathbf{x})$$

and

$$\Pr(Y = r|\mathbf{x}, \mathbf{w}) = \pi(\mathbf{x}) \frac{\Gamma(r + k)}{\Gamma(r + 1)\Gamma(k)} \times \left[1 - \left\{1 + \frac{\lambda(\mathbf{w})}{k}\right\}^{-k}\right]^{-1} \left\{\frac{\lambda(\mathbf{w})}{k}\right\}^{r} \left\{1 + \frac{\lambda(\mathbf{w})}{k}\right\}^{-(r+k)}$$

for $r = 1, 2, 3, \ldots$, where $\pi(\mathbf{x}) = \Pr(Z = 1|\mathbf{x})$ is the probability that the species is present, and $\lambda(\mathbf{w})$ is the mean of the untruncated negative binomial distribution, the mean of the truncated version being

$$E(Y|Z = 1) = \lambda(\mathbf{w}) \left\{ 1 - \left(1 + \frac{\lambda(\mathbf{w})}{k} \right)^{-k} \right\}^{-1}$$

(Johnson, Kemp, and Kotz 2005, chap. 5).

Using a logistic link and log link for $\pi(\mathbf{x})$ and $\lambda(\mathbf{w})$, respectively, we have

$$\log \left\{ \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} \right\} = \mathbf{x}^{\mathrm{T}} \boldsymbol{\beta},$$

and

$$\log\{\lambda(\mathbf{w})\} = \mathbf{w}^{\mathrm{T}} \boldsymbol{\theta},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ and $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_q)^{\mathrm{T}}$ are vectors of unknown parameters.

Suppose we have abundances y_i, of which m are positive ($i = 1, 2, \ldots, n$). Let the corresponding values for the covariates be denoted $\mathbf{x}_i = (1\ x_{1i}, \ldots, x_{pi})^{\mathrm{T}}$ and $\mathbf{w}_i = (1\ w_{1i}, \ldots, w_{qi})^{\mathrm{T}}$ ($i = 1, 2, \ldots, n$). The log-likelihood for the data is given by

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\theta}, k) = & \sum_{y_i = 0} \log(1 - \pi_i) + \sum_{y_i > 0} \log \pi_i + m\{k \log k - \log \Gamma(k)\} \\ & + \sum_{y_i > 0} \Bigg[\log \Gamma(y_i + k) - \log \Gamma(y_i + 1) + y_i \log \lambda_i \\ & - (y_i + k) \log(k + \lambda_i) - \log \left\{ 1 - \left(1 + \frac{\lambda_i}{k} \right)^{-k} \right\} \Bigg] \end{aligned} \tag{3.1}$$

where $k > 0$, and we have

$$\pi_i = \pi(\mathbf{x}_i) = \{1 + \exp(-\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})\}^{-1},$$

and

$$\lambda_i = \lambda(\mathbf{w}_i) = \exp(\mathbf{w}_i^{\mathrm{T}} \boldsymbol{\theta}).$$

Note that the "overdispersion" parameter k is assumed to not depend on any of the covariates.

Suppose we wish to estimate the expected abundance for particular values of the covariates, $\mathbf{x}_0 = (1\ x_{10}, \ldots, x_{p0})^{\mathrm{T}}$ and $\mathbf{w}_0 = (1\ w_{10}, \ldots, w_{q0})^{\mathrm{T}}$. This is given by

$$\begin{aligned} \mu_0 &= E(Y|\mathbf{x}_0, \mathbf{w}_0) \\ &= \pi_0 \lambda_0 \left\{ 1 - \left(1 + \frac{\lambda_0}{k} \right)^{-k} \right\}^{-1} \end{aligned} \tag{3.2}$$

where, for ease of notation, we use π_0 and λ_0 to denote $\pi(\mathbf{x}_0)$ and $\lambda(\mathbf{w}_0)$, respectively. The maximum likelihood estimate of μ_0 is given by

$$\hat{\mu}_0 = \hat{\pi}_0 \hat{\lambda}_0 \left\{ 1 - \left(1 + \frac{\hat{\lambda}_0}{\hat{k}} \right)^{-\hat{k}} \right\}^{-1},$$

where

$$\hat{\pi}_0 = \left\{1 + \exp(-\mathbf{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}})\right\}^{-1}$$

and

$$\hat{\lambda}_0 = \exp(\mathbf{w}_0^{\mathrm{T}}\hat{\boldsymbol{\theta}}),$$

and $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{k}$ are the maximum likelihood estimates of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and k, respectively.

Calculation of a profile likelihood interval for some quantity of interest involves maximizing the log-likelihood under the constraint that that quantity takes a fixed value v, say. The profile log-likelihood is the maximized likelihood as a function of v. The quantity of interest here is the unconditional mean (3.2), and so the relevant constrained maximization is equivalent to unconstrained maximization of:

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}, k) - \gamma\pi_0\lambda_0\left\{1 - \left(1 + \frac{\lambda_0}{k}\right)^{-k}\right\}^{-1}, \tag{3.3}$$

where γ is a Lagrange multiplier (Khuri 1993, p. 287). We maximize (3.3) for a range of values of γ, and the corresponding values of μ_0 are given by substituting the resulting estimates of π_0, λ_0, and k into (3.2). Thus, varying γ amounts to varying μ_0 with $\gamma = 0$ corresponding to the maximum likelihood estimate $\hat{\mu}_0$.

Note that use of a Lagrange multiplier is equivalent to writing one of the original parameters in (3.1) in terms of μ_0 and working with the resulting log-likelihood as a function of the remaining original parameters and μ_0. This is true even when such a substitution cannot be performed explicitly. Note also that in performing the maximization of (3.3) numerically, it is a good idea to replace

$$\pi_0\lambda_0\left\{1 - \left(1 + \frac{\lambda_0}{k}\right)^{-k}\right\}^{-1}$$

by some bounded monotone function of it, such as $x/(1 + x)$, to prevent this component becoming too large.

The profile log-likelihood for μ_0 is defined as

$$l_p(\mu_0) = l\{\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}}, \tilde{k}\},$$

where $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\theta}}$, and $\tilde{k}$ are the values of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and k that maximize (3.3) for a given value of μ_0 (or equivalently γ). Note that the values of $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\theta}}$, and $\tilde{k}$ will depend on μ_0, and that we have suppressed this dependence for ease of notation. Large-sample likelihood theory (Venzon and Moolgavkar 1988) tells us that

$$w(\mu_0) = 2\{l_p(\hat{\mu}_0) - l_p(\mu_0)\}$$

will have a distribution which is approximately χ_1^2. This leads to a 95% profile likelihood confidence interval being defined as the two values of μ_0 that satisfy $w(\mu_0) = 3.8415$, with 3.8415 being the 95th percentile of the χ_1^2 distribution.

The estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\theta}}$, $\hat{k}$, $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\theta}}$, and $\tilde{k}$ are all obtained using numerical methods. For the example in the next section, we performed the calculations in MATLAB, using the Nelder-Mead simplex algorithm for numerical optimization.

Table 2. Parameter estimates for the two components of the conditional model, together with 95% profile likelihood confidence intervals.

Model	Parameter	Estimate	CI
Probability of presence	β_0	2.18	(1.17, 3.41)
log-likelihood = –55.94	β_1	–0.45	(–0.85, –0.10)
Conditional abundance	θ_0	4.20	(3.78, 4.65)
log-likelihood = –366.71	θ_1	–0.18	(–0.36, –0.0077)
	k	1.01	(0.67, 1.43)

4. EXAMPLE

As we focus on a single covariate in our example, we have $\mathbf{x} = \mathbf{w}$, $p = q = 1$, and there are five parameters: β_0, β_1, θ_0, θ_1, and k. Table 2 shows the maximum likelihood estimates for these parameters, together with 95% profile likelihood confidence intervals.

Figure 2 shows the profile log-likelihood as a function of the expected abundance of *Ecklonia* when the abundance of *Evechinus* is zero. The "cut" for a 95% confidence interval is also shown, together with a quadratic approximation to this profile, corresponding to a Wald interval. The asymmetry about the maximum likelihood estimate of the profile log-

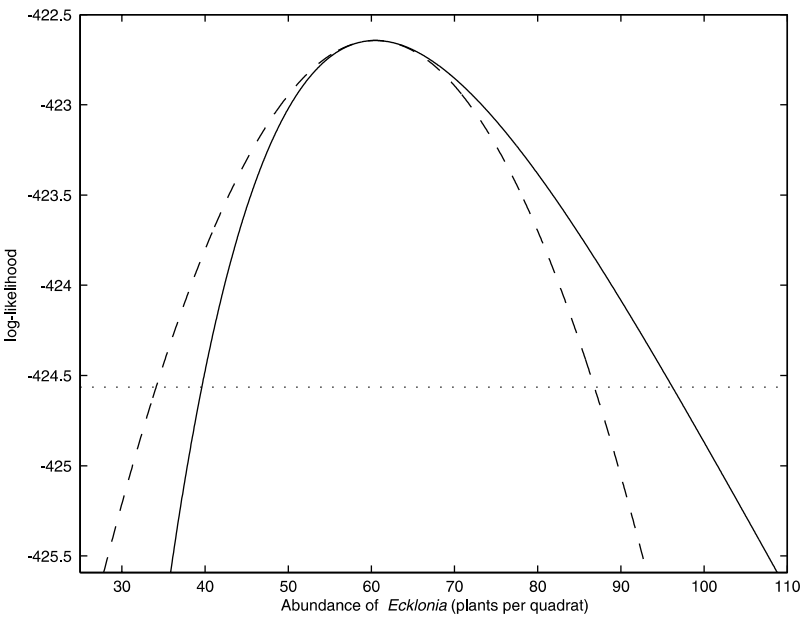

Figure 2. Profile log-likelihood for expected abundance of *Ecklonia* when abundance of *Evechinus* is zero (——), with a quadratic approximation (– – –) and the "cut" for a 95% confidence interval (· · ·).

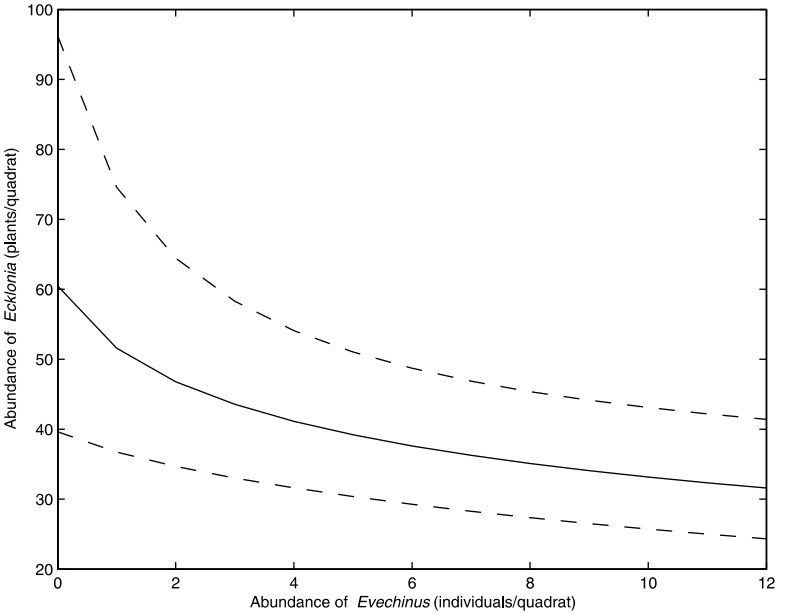

Figure 3. Estimates and 95% profile likelihood confidence intervals for the abundance of *Ecklonia*, over a range of values for the abundance of *Evechinus*.

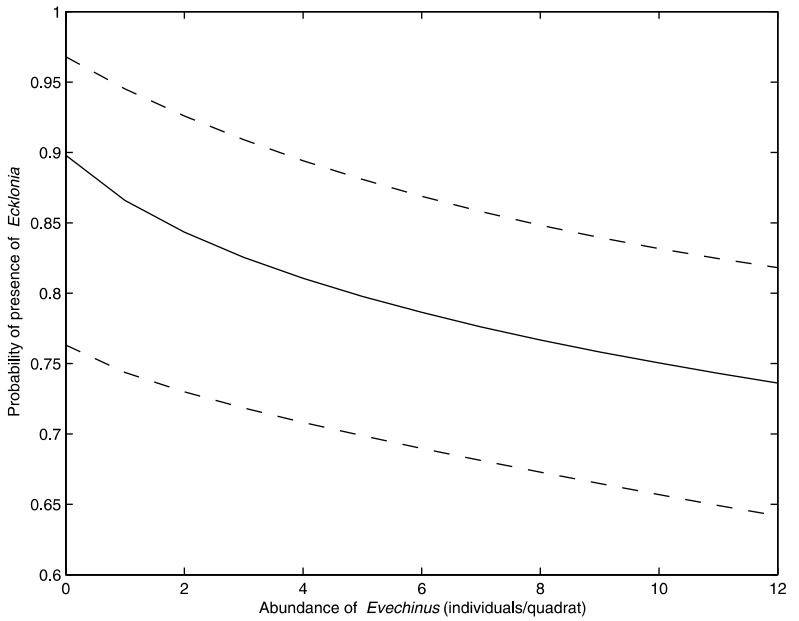

Figure 4. Estimates and 95% profile likelihood confidence intervals for the probability of presence of *Ecklonia*, over a range of values for the abundance of *Evechinus*.

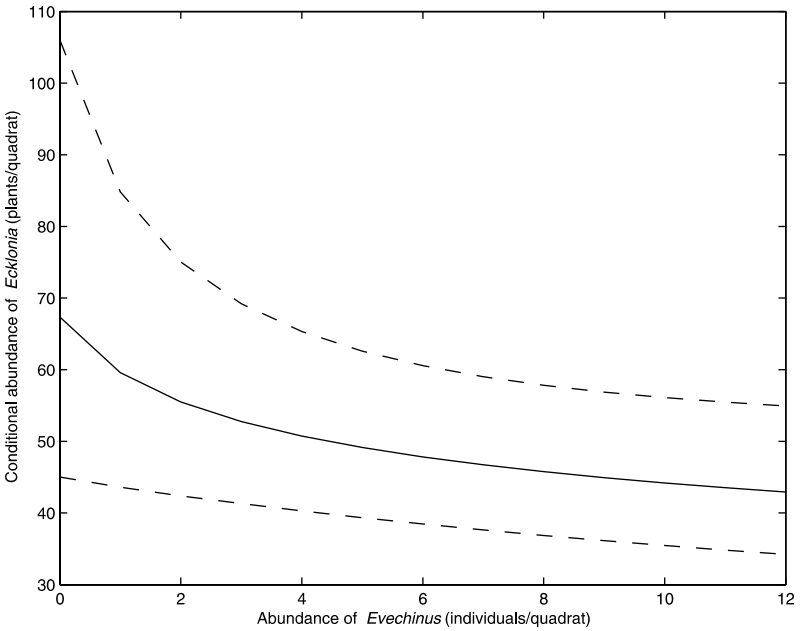

Figure 5. Estimates and 95% profile likelihood confidence intervals for the conditional abundance of *Ecklonia*, over a range of values for the abundance of *Evechinus*.

likelihood interval shows that there is less uncertainty in the estimation below the maximum likelihood estimate than above it.

Figure 3 shows estimates and 95% profile likelihood confidence intervals for the expected abundance of *Ecklonia* over a range of values for the abundance of *Evechinus*. The confidence intervals clearly aid interpretation of the effect of *Evechinus* abundance on *Ecklonia*. In addition, the asymmetry achieved using a profile likelihood interval is intuitively sensible.

The generalized Pearson statistic (sum of squared standardized residuals) is 99.3 on 98 df, suggesting an adequate fit. The circled observation in Figure 1 has the largest standardized residual, with a tail probability of 0.0014. This might be considered an outlier, but it is not altogether extreme in a sample of 103 observations. As this observation also appears influential, we checked the effect of removing it from the analysis. The difference on the fitted model is not great (the log-likelihood increased by 0.76), with the estimates of θ_0, θ_1, and k (Table 2) changing by $+0.11$, -0.08, and $+0.12$, respectively (the estimates of β_0 and β_1 are unaffected). The next largest standardized residual corresponds to a tail probability of 0.026, which would seem unremarkable in a sample of 103 observations.

Although they are not the focus of this article, we also calculated estimates and 95% profile likelihood confidence intervals for the probability of presence of *Ecklonia* (Figure 4) and the conditional abundance of *Ecklonia* (Figure 5), respectively. These graphs complement Figure 2, in that they aid understanding as to how the two components of abundance of *Ecklonia* are influenced by abundance of *Evechinus*.

Finally, we note that the approach we present here is quite general and could be applied,

in principle, to any situation where we wish to calculate a confidence interval for the expected value from a statistical model. It is likely to be of most value when a symmetric confidence interval might provide a poor reflection of the uncertainty in the estimate of the expected value, due to asymmetry of the profile log-likelihood function.

[Received July 2006. Revised February 2007.]

REFERENCES

Barry, S. C., and Welsh, A. H. (2002), "Generalized Additive Modelling and Zero Inflated Count Data," *Ecological Modelling*, 157, 179–188.

Brown, L. D., Cai, T., and DasGupta, A. (2003), "Interval Estimation in Exponential Families," *Statistica Sinica*, 13, 19–49.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman and Hall.

Dobbie, M. J., and Welsh, A. H. (2001), "Modelling Correlated Zero-Inflated Count Data," *Australian and New Zealand Journal of Statistics*, 43, 431–444.

Fletcher, D. J., MacKenzie, D. I., and Villouta, E. (2005), "Modelling Skewed Data with Many Zeros: A Simple Approach Combining Ordinary and Logistic Regression," *Environmental and Ecological Statistics*, 12, 45–54.

Johnson, N.L., Kemp, A.W., and Kotz, S. (2005), *Univariate Discrete Distributions* (3rd ed.), New York: Wiley.

Khuri, A.I. (1993), *Advanced Calculus with Applications in Statistics*, New York: Wiley.

Lambert, D. (1992), "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing," *Technometrics*, 34, 1–14.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., and Possingham, H. P. (2005), "Zero Tolerance Ecology: Improving Ecological Inference by Modelling the Source of Zero Observations," *Ecology Letters*, 8, 1235–1246.

Stephenson, G. (1961), *Mathematical Methods for Science Students*, London: Longmans.

Venzon, D. J., and Moolgavkar, S. H. (1988), "A Method for Computing Profile-Likelihood-Based Confidence Intervals," *Applied Statistics*, 37, 87–94.

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.