Implementing a generic method for bias correction in statistical models using random effects, with spatial and population dynamics examples

James T. Thorson [a,*], Kasper Kristensen [b]

[a] Fisheries Resource Analysis and Monitoring Division, Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 2725 Montlake Blvd. E, Seattle, WA 98112, United States
[b] Technical University of Denmark, Charlottenlund Slot Jægersborg Allé 1, 2920 Charlottenlund, Denmark

ABSTRACT

Statistical models play an important role in fisheries science when reconciling ecological theory with available data for wild populations or experimental studies. Ecological models increasingly include both fixed and random effects, and are often estimated using maximum likelihood techniques. Quantities of biological or management interest ("derived quantities") are then often calculated as nonlinear functions of fixed and random effect estimates. However, the conventional "plug-in" estimator for a derived quantity in a maximum likelihood mixed-effects model will be biased whenever the estimator is calculated as a nonlinear function of random effects. We therefore describe and evaluate a new "epsilon" estimator as a generic bias-correction estimator for derived quantities. We use simulated data to compare the epsilon-method with an existing bias-correction algorithm for estimating recruitment in four configurations of an age-structured population dynamics model. This simulation experiment shows that the epsilon-method and the existing bias-correction method perform equally well in data-rich contexts, but the epsilon-method is slightly less biased in data-poor contexts. We then apply the epsilon-method to a spatial regression model when estimating an index of population abundance, and compare results with an alternative bias-correction algorithm that involves Markov-chain Monte Carlo sampling. This example shows that the epsilon-method leads to a biologically significant difference in estimates of average abundance relative to the conventional plug-in estimator, and also gives essentially identical estimates to a sample-based bias-correction estimator. The epsilon-method has been implemented by us as a generic option in the open-source Template Model Builder software, and could be adapted within other mixed-effects modeling tools such as Automatic Differentiation Model Builder for random effects. It therefore has potential to improve estimation performance for mixed-effects models throughout fisheries science.

Published by Elsevier B.V.

1. Introduction

Statistical models for the dynamics of wild populations are designed to reconcile available data with ecological theory, and are often used to make predictions about future, historical, or otherwise unobserved process or events (Hilborn and Mangel, 1997). For example, population models are commonly used when evaluating trade-offs arising from management decisions, e.g., determining what level of fishery catch is consistent with long-term management objectives of fishery stakeholders (Walters and Martell, 2004). Population models are also used to improve insight regarding environmental trends at large spatial scales, e.g., detecting long-term trends in abundance for populations of birds (Schaub et al., 2007).

Parameters in population dynamics models are often estimated by identifying the values of parameters that maximize the "likelihood function". The likelihood function is defined as the probability that the observed data would arise given a hypothesized model and proposed set of parameter values. Modern population models increasingly include a mix of random and fixed effects, where random effects are assumed to arise from a random process and generally account for non-independence in available data (Thorson and Minto, 2015). In particular, random effects may be used to approximate spatial variation in population density and productivity (Kristensen et al., 2014; Thorson et al., 2015b), variation in growth rates among individuals (Shelton

* Corresponding author. Fax: +1 206 860 6792.
  E-mail address: James.Thorson@noaa.gov (J.T. Thorson).

et al., 2013), and variation over time in survival rates (Minto et al., 2014). Random effects can also be used to approximate more complicated stochastic processes, e.g., the unknown function representing the relationship between population abundance and population growth or recruitment (Munch et al., 2005; Thorson et al., 2014b).

Ecologists are often interested in predictions arising from population models, and these predicted quantities may be a function of random effects. As one example, population abundance could be calculated as the sum of density at multiple sites, where the logarithm of density $\log(d_i)$ at each site i is modeled as a normally-distributed random effect, i.e., $\log(d_i) = \gamma_i$, where $\gamma_i \sim N\left(\mu, \sigma^2\right)$ and μ and σ^2 are the mean and variation of log-density among sites. In this example, the random effect γ_i represents log-density at each site, and γ_i is transformed via a nonlinear function (exponentiated) when calculating the derived variable of interest (site-specific density). An ecologist might be interested in different "levels" of this hierarchical model, e.g., either the average density among sites μ or the predicted density at any individual site d_i.

Whenever a random variable is transformed by a nonlinear function, the mean and variance of the variable are also transformed. This is handled smoothly when estimating parameters using Bayesian methods (i.e., given samples from the posterior probability of parameters) because the value of the derived quantity can be calculated for each realization of the posterior distribution. However, in maximum likelihood estimation, the derived quantity is frequently calculated using a "plug-in" estimator, i.e., $\hat{d}_i = \exp\left(\hat{\gamma}_i\right)$ where $\hat{\gamma}_i$ is the estimate of log-density at site i, and $\hat{d}_i$ is the plug-in estimator for density. Given an unbiased estimate of log-density $\hat{\gamma}_i$ in this example, the plug-in estimator of density $\hat{d}_i = \exp\left(\hat{\gamma}_i\right)$ will be biased because the function $\exp\left(\hat{\gamma}_i\right)$ is nonlinear. Previous research has developed methods to calculate the expected value of a derived quantity for specific fisheries models. For example, Methot and Taylor (2011) developed an algorithm to approximate the expected value of recruitment for marine populations, because recruitment is often calculated as a nonlinear function of a random effect representing temporal variation in survival of juvenile fish. Methot and Taylor (2011) note that their algorithm is not necessary for a Bayesian model of recruitment, because the expected value can be calculated for any nonlinear function given a set of samples from the posterior distribution of parameters. However, most stock assessment models used for tactical management in the United States and elsewhere continue to use maximum likelihood estimation methods, so the Methot and Taylor (2011) algorithm remains important for many stock assessment models.

Here we demonstrate a novel approach for approximating the expected value of a quantity derived from random effects in a mixed-effects model. We first explain the theory of derived quantities in more detail, and explain the proposed "epsilon" method. We then provide a simulation experiment comparing the epsilon method with the previous Methot and Taylor (2011) algorithm for the case of recruitment in population models for marine populations, using four levels of data availability. We conclude by applying the epsilon method when estimating expected population abundance using a spatial regression model and a real-world application. This case study shows that the bias-corrected estimate from the epsilon method (and its estimated variance) are similar to estimates from an alternative, sample-based bias-correction estimator. We have implemented the epsilon method in the new Template Model Builder software (publicly available at https://github.com/kaskr/adcomp), and it can be applied off-the-shelf to other mixed-effects models.

2. Methods

2.1. Defining bias in maximum likelihood estimates of mixed-effects models

Fixed effects are parameters that are estimated by identifying which values maximize the likelihood function, defined as the probability (likelihood) of obtaining the observed data given a hypothesized value and proposed values for the fixed effects. By contrast, random effects are treated as random variables, and therefore are represented by a distribution of possible values where that distribution generally has a finite mean and variance. For mixed-effects models, the marginal likelihood function is calculated by integrating across all possible values of the random effects. Random effects are included in statistical models to account for situations in which available data are not statistically independent (e.g., variation over time, space, or among individuals), and mixed-effects models represent a generic approach for calculating the covariation among data that arises from a hypothesized model (Thorson and Minto, 2015). Values for random effects in maximum likelihood models can be predicted after fixed effects have been estimated (we note that other authors sometimes refer to the task of predicting random effects based on estimated fixed effects as "estimating" the random effects). Values for random effects are generally predicted via "empirical Bayes" (de Valpine, 2009), i.e., by plugging in the maximum likelihood estimate of fixed effects and then setting random effects to the mode or mean of their conditional distribution. However, it is possible to estimate fixed effects by maximizing the likelihood function without ever predicting the value of random effects. In the following, we are concerned with providing unbiased predictions of quantities that are calculated as a nonlinear function of random effects, and this in no way affects the estimates of fixed effects that arise from maximum likelihood estimation.

Maximum likelihood estimates of fixed effects are generally "consistent", i.e., will converge to their true value given that samples sizes increase indefinitely (given reasonable conditions and that the model is correct; see Gelman et al. (2003, pg. 107–111)). By contrast, Empirical Bayes predictions of random effects are not generally consistent. Instead, additional data will often result in an increase in the number of random effects that must be predicted, thus maintaining a constant ratio of data per random effect.

Despite providing a consistent estimate of fixed effects, maximum likelihood estimates may be biased. An estimator $\hat{\theta}$ of parameter θ is said to be biased if $E\left[\hat{\theta}\right] - \theta \neq 0$ for a given sample size, and maximum likelihood estimates are often biased given small sample size. Bias-correction methods generally exist, and can be used to calculate a "bias-corrected" estimator $\hat{\theta}_{BC}$ that is unbiased for all sample sizes (i.e., where $E[\hat{\theta}_{BC}] - \theta = 0$ regardless of sample size). However, bias-corrected estimators generally increase the total estimation variance, i.e., $\text{Var}\left[\hat{\theta}_{BC} - \theta\right] > \text{Var}\left[\hat{\theta}_{PI} - \theta\right]$ (where the left-hand side is the variance of the bias-corrected estimator, and the right-hand side is the variance of the plug-in estimator $\hat{\theta}_{PI}$ without bias correction). There is therefore a tradeoff between bias and variance when deciding whether to use bias-correction methods when estimating parameters via maximum likelihood. This definition of bias and bias-correction applies to both estimates of fixed-effects (via the likelihood function), and predictions of random effects (conditional on estimated fixed effects). In the following, we are specifically concerned with bias-correction for predictions of random effects, and a general treatment of bias-correction for fixed-effects in nonlinear mixed-effects models is not addressed here.