Ecological Indicators 85 (2018) 502–508

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Research paper

A study on the effects of unbalanced data when fitting logistic regression models in ecology

Christian Salas-Eljatib[a,*], Andres Fuentes-Ramirez[a], Timothy G. Gregoire[b], Adison Altamirano[c], Valeska Yaitul[a]

[a] Laboratorio de Biometría, Departamento de Ciencias Forestales, Universidad de La Frontera, Temuco, Chile
[b] School of Forestry and Environmental Studies, Yale University, New Haven, CT 065111, USA
[c] Laboratorio de Ecología del Paisaje Forestal, Departamento de Ciencias Forestales, Universidad de La Frontera, Temuco, Chile

ARTICLE INFO

Keywords:
Statistical inference
Model prediction
Logit model
Binary variable
Bias
Precision

ABSTRACT

Binary variables have two possible outcomes: occurrence or non-occurrence of an event (usually with 1 and 0 values, respectively). Binary data are common in ecology, including studies of presence/absence, alive/dead, and change/no-change. Logistic regression analysis has been widely used to model binary response variables. Unbalanced data (i.e., an extremely larger proportion of zeros than ones) are often found across a variety of ecological datasets. Sometimes the data are balanced (i.e., same amount of zeros and ones) before fitting the model, however, the statistical implications of balancing (or not) the data remain unclear. We assessed the statistical effects of balancing data when fitting a logistic regression model by studying both its statistical properties of the estimated parameters and its predictive capabilities. We used a base forest-mortality model as reference, and by using stochastic simulations representing different configurations of 0/1 data in a sample (unbalanced data scenarios), we fitted the logistic regression model by maximum likelihood. For each scenario we computed the bias and variance of the estimated parameters and several prediction indexes. We found that the variability of the estimated parameters is affected, with the balanced-data scenario having the lowest variability, thus, affecting the statistical inference as well. Furthermore, the prediction capabilities of the model are altered by balancing the data, with the balanced-data scenario having the better sensitivity/specificity ratio. Balancing, or not, the data to be used for fitting a logistic regression models may affect the conclusion that can arise from the fitted model and its subsequent applications.

1. Introduction

Data of occurrence/non-occurrence of a phenomenon of interest are vastly found across several disciplines (Alberini, 1995; Arana and Leon, 2005; Bell et al., 1994). This type of variable is known as binary or dichotomous, and it represents whether an event occurs or not. This event is represented by the random variable Y, and we usually record occurrence by $Y = 1$ and non-occurrence by $Y = 0$. In ecology, binary variables arise when studying the presence of a species in a geographic area (Bastin and Thomas, 1999; Phillips and Elith, 2013; Hastie and Fithian, 2013) or the occurrence of mortality at the tree or forest level (Davies, 2001; Wunder et al., 2008; Chao et al., 2009; Young et al., 2017). Meanwhile in landscape ecology, binary variables are used to represent the occurrence of fire within a given area (Bigler et al., 2005; Mermoz et al., 2005; Dickson et al., 2006; Vega-García and Chuvieco, 2006; Palma et al., 2007; Bradstock et al., 2010); deforestation (Wilson et al., 2005; Schulz et al., 2011; Kumar et al., 2014; Hu et al., 2014); and in general the change from one land use category to another (Seto and Kaufmann, 2005; Leyk and Zimmermann, 2007; Lander et al., 2011).

Logistic regression analysis is the most frequently used modelling approach for analyzing binary response variables. If we need to model a binary variable, to statistically relate it to predictor variable(s) or covariate(s), one of the most used approaches for pursuing this task in ecology is to use logistic regression models (Warton and Hui, 2011). These models belong to group of the generalized linear models (GLM). In a GLM, three compartments must be specified (Lindsey, 1997; Schabenberger and Pierce, 2002): a random component, a systematic component, and a link function. A logistic regression model uses: a binomial probability density function as the random component; a linear predictor function $\mathbf{X}'\boldsymbol{\beta}$ (where $\mathbf{X}$ is a matrix with the covariates and $\boldsymbol{\beta}$ is a vector with the parameters or coefficients) as the systematic component; and a logistic equation as the link function. One of the key advantages of using logistic regression models in ecology is that the

* Corresponding author. Tel.: +56 45 2325652.
E-mail address: christia.salas@ufrontera.cl (C. Salas-Eljatib).

http://dx.doi.org/10.1016/j.ecolind.2017.10.030
Received 7 April 2017; Received in revised form 9 August 2017; Accepted 16 October 2017

probability of the binary response variable is directly modelled, thereby accounting explicitly for the random nature of the phenomenon of interest.

In many applications when dealing with binary data in ecology, it happens that the number of observations with ones ($Y = 1$) is much smaller than the number of observations with zeros ($Y = 0$) or vice versa. We simply term this situation as *unbalanced data*, but other terms have been also used for this situation, including disproportionate sampling (Maddala, 1992) or rarity events (King and Zeng, 2001). Based on our review of scientific applications of logistic regression to model ecological phenomena, the proportion of zeros in datasets ranges between 80% and 95%. Therefore, having balanced data (i.e., equal numbers of observations of zeros and ones) is more the exception than the rule.

Both unbalanced and balanced data have been used for fitting logistic regression models. In ecological studies, some researchers have adopted the practice of balancing the data before carrying out the analyses (e.g., Vega-García et al., 1995; Vega-García et al., 1999; Lloret et al., 2002; Brook and Bowman, 2006; Vega-García and Chuvieco, 2006; Jones et al., 2010; Rueda, 2010). Balancing data means to select, by some rule (usually at random), the same amount of observations with ones and zeros from the originally available dataset. Therefore, a balanced dataset or balanced sample is created, where a 50–50% proportion of zero and one values is met. After the balanced dataset is built, the logistic regression model is fitted (i.e., its parameters are estimated) by maximum likelihood (ML). An example of this practice in ecological applications is the option for balancing data before fitting a logit model when conducting analyses of land use changes in the software IDRISI (Eastman, 2006). On the other hand, it is important to point out that unbalanced data have been also used in ecological studies (Wilson et al., 2005; Echeverria et al., 2008; Kumar et al., 2014; Young et al., 2017). Therefore, unbalanced data in applied ecological studies has been considered as not having important effects into the models being fitted. Moreover, to date, no studies have addressed the effect of balancing data when fitting logistic regression models in ecological analyses, and just a handful have explored some statistical implications in ecological applications (Qi and Wu, 1996; Wu et al., 1997; Cailleret et al., 2016).

The applied statistical implications of unbalanced data in logistic regression are not well described nor realized for applied researchers. Although balancing the data seems to be an accepted practice, the reasons that justify its use are not well explained. The most immediate effect of balancing the data is to greatly reduce the sample size available for fitting purposes, therefore decreasing the precision with which the parameters of the model are estimated. Among the statistical studies on logistic regression and unbalanced data, we highlight the following: Schaefer (1983) and Scott and Wild (1986) pointed out that the maximum likelihood estimates (MLE) of a logit model are biased only for small sample sizes. On the other hand, Xie and Manski (1989) stated that unbalanced data only affect the intercept parameter of a logit model, specifically being biased estimated according to Maddala (1992). King and Zeng (2001), advocated that all the MLE of the logit parameters are biased. Schaefer (1983) and Firth (1993) proposed correction for the bias of the MLE of the logistic regression model parameters. McPherson et al. (2004) conducted one of the few related analysis when fitting presence-absence species distribution models in ecology, but only focusing in the prediction capabilities of the fitted models. Maggini et al. (2006) assessed the effect of weighting absences when modelling forest communities by generalized additive models. Recently, Komori et al. (2016) indicated that logistic regression suffer poor predictive performance, and proposed an alternative model to improve predictive performance. Komori et al. (2016) approach involves to add a new parameter to the original structure of a logistic regression model, and fitted it in a mixed-effects modelling framework, therefore their approaches becomes a different type of statistical model. From above, we can infer that: (a) most of the statistical studies on logistic regression and unbalanced data have focus on the bias of the MLE parameters (a topic that has been rarely taking into account in ecological applications); (b) much less attention has been put into the prediction performance; and (c) no study has dealt with the effects of unbalanced data in the variance of the MLE parameters.

In this study we aimed at assessing the effect of using unbalanced data when fitting logistic regression models by analyzing both the statistical properties (i.e., bias and variance) of the estimated parameters and the predictive capabilities of the fitted model.

2. Methods

2.1. Base model

The binary variable (Y) is the occurrence of a phenomenon of interest, where $Y = 1$ denotes occurrence and $Y = 0$ otherwise. In a modelling framework, we seek to model the probability of the response variable being $Y = 1$, given the values of the predictor variables, this is $\Pr(Y = 1|\mathbf{X})$, that we can more easily represent by $\pi_{\mathbf{y}|\mathbf{X}}$.

In our analysis we used a logistic regression equation with five predictor variables, as a base model for carrying out our analysis. This model served as a reference for assessing the statistical effects of unbalanced data on fitting logistic regression models. The binary variable of forest mortality occurrence (Y), given the analyses of Young et al. (2017) in the state of California, USA, is modeled as a function of climate and biotic variables, as follows:

$$\ln\left[\frac{\pi_{y_i|x_i}}{1-\pi_{y_i|x_i}}\right] = \text{logit}[Y_i = 1] = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2i} + \beta_3 X_{3_i} + \beta_4 X_{4_i} + \beta_5 X_{5_i}, \quad (1)$$

where Y_i is the occurrence of forest mortality (i.e., 1 for occurrence, 0 for non-occurrence) at the ith pixel), meanwhile the predictor variables X_{1i}, X_{2i}, X_{3i}, X_{4i}, and X_{5i} represent the: mean climatic water deficit (CWD) or simply $Defnorm_i$; basal area of live trees (BA_i); BA_i^2; CWD anomaly ($Defz0_i$); and $Defnorm_i \times BA_i$ for the ith pixel, respectively. We have used the nomenclature for the variables as in the study of Young et al. (2017) and only the available data for year 2012. Notice that we could more easily represent model (1) as:

$$\ln\left[\frac{\pi_{\mathbf{y}|\mathbf{X}}}{1-\pi_{\mathbf{y}|\mathbf{X}}}\right] = \text{logit}[\mathbf{y}{=}1] = \mathbf{X}'\boldsymbol{\beta}, \quad (2)$$

where $\mathbf{y}$ is the vector with the binary variable, $\mathbf{X}$ is the matrix with the predictor variables (and a first column of 1), and $\boldsymbol{\beta}$ is the vector of parameters $[\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}, \widehat{\beta_3}, \widehat{\beta_4}, \widehat{\beta_5}]$.

In the sequel, we shall use Eq. (2) as the mean function in various scenarios of unbalanced data. It is important to point out that we are not interested in finding the best model, but rather on studying the effects of using several unbalanced data scenarios on a reference model. Furthermore, we want to remark that we are not pursuing to assess different alternative statistical models for unbalanced data (e.g. as in, Warton and Hui, 2011; Hastie and Fithian, 2013). We also want to mention that the zero-inflated models are those focusing on modelling count variables (Schabenberger and Pierce, 2002; Zuur et al., 2010), such as the prediction of the amount of tree mortality (e.g., Affleck, 2006). These models are not part of our study, since we are dealing with modelling a binomial variable.

2.2. Unbalanced data scenarios

We use data of forest mortality occurrence from Young et al. (2017), in California during 2012 as our population, containing 11763 total observations (N), with 2985 cases of mortaltity occurrence (N_1) and 8778 cases of non-occurrence (N_0). In order to assess the effects of unbalanced data on the statistical properties of the logit model (Eq.