Journal of Asia-Pacific Biodiversity 9 (2016) 392–395

HOSTED BY

Contents lists available at ScienceDirect

Journal of Asia-Pacific Biodiversity

journal homepage: http://www.elsevier.com/locate/japb

Short communication

A maximum likelihood model for fitting power functions with data uncertainty: A case study on the relationship between body lengths and masses for Sciuridae species worldwide

Youhua Chen*

Department of Renewable Resources, University of Alberta, Edmonton, AB, Canada

A R T I C L E I N F O

Article history:
Received 24 December 2015
Received in revised form
5 March 2016
Accepted 17 March 2016
Available online 31 March 2016

Keywords:
allometry
maximum likelihood
numerical optimization
power law
stochasticity

A B S T R A C T

In this report, a maximum likelihood model is developed to incorporate data uncertainty in response and explanatory variables when fitting power-law bivariate relationships in ecology and evolution. This simple likelihood model is applied to an empirical data set related to the allometric relationship between body mass and length of Sciuridae species worldwide. The results show that the values of parameters estimated by the proposed likelihood model are substantially different from those fitted by the nonlinear least-of-square (NLOS) method. Accordingly, the power-law models fitted by both methods have different curvilinear shapes. These discrepancies are caused by the integration of measurement errors in the proposed likelihood model, in which NLOS method fails to do. Because the current likelihood model and the NLOS method can show different results, the inclusion of measurement errors may offer new insights into the interpretation of scaling or power laws in ecology and evolution.

Introduction

Bivariate relationships have been widely studied in the fields of ecology and evolution. A suite of bivariate relationships has been widely documented in previous literature, including allometric relationship (Manaster and Manaster 1975; White 2011), species-area relationship (Connor and McCoy 1979; Rosenzweig 1995; Solmos and Lele 2012), etc. Among the alternative statistical models in bivariate analysis, the least-of-square fitting technique (Leonard 2011) is one of the most appropriate methods applied for investigating allometric relationship (White 2011) or species-area relationship (Connor and McCoy 1979; Triantis et al 2012).

Conventionally statistical methods typically do not take into account the issue of data uncertainty. In these methods, the average is taken when one object is measured multiple times. These averages over different objects are then used for ordinary least square fitting. However, this averaging practice might result in loss of a lot of information inherited in the raw data, because the average of different individual measures cannot reflect the dispersion of the data and the contribution of individual variation (Felsenstein 2008; Ives et al 2007; Revell and Reynolds 2012; Violle et al 2012). As such, when one wants to better quantify exact slope values in power-law bivariate models, the influence of data uncertainty indicated by the standard deviation of the data should be not neglected. Therefore, it is necessary to develop new statistical methods to cope with data uncertainty issue for fitting bivariate relationships. Accordingly, the central objective of this work is to develop a simple maximum likelihood model to estimate the associated parameters in bivariate models by considering the uncertainty of the raw data.

Materials and methods

A maximum likelihood model is developed for measuring uncertainties in both response and explanatory variables in bivariate regression models.

Similar to a previous study (Ma et al 2013), the maximum likelihood model is formulated as follows: assume that in the empirical data set each data point (x, y) is supplied with measurement errors in both explanatory and response variables as (dx, dy). The bivariate function that is required to fit is $Y = f(X)$, where (X, Y) represents any point on the fitted curve.

* Corresponding author.
E-mail address: haydi@126.com.
Peer review under responsibility of National Science Museum of Korea (NSMK) and Korea National Arboretum (KNA).

http://dx.doi.org/10.1016/j.japb.2016.03.015
pISSN2287-884X eISSN2287-9544

In practice, finding a fitted point on the fitted curve closest to the focused empirical point (x, y) has the highest explanatory power. Then, it is necessary to measure the following quantity to minimize the difference

$$\begin{aligned} y - f(x) \\ &= (y - Y) + [Y - f(x)] \\ &= (y - Y) + [f(X) - f(x)] \\ &= (y - Y) + f'(x)(X - x) \end{aligned} \quad (1)$$

As such, the variance for the left-side quantity is given by (Ma et al 2013)

$$\begin{aligned} \mathrm{Var}[y - f(x)] \\ &= \mathrm{Var}(y - Y) + f'(x)^2 \mathrm{Var}[(X - x)] \\ &= dy^2 + f'(x)^2 dx^2 \end{aligned} \quad (2)$$

Consequently, the likelihood model for many empirical points can be formulated as follows:

$$\mathrm{Likelihood}[f(x)] = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi\left[dy_k^2 + f'(x_k)^2 dx_k^2\right]}} \exp\left(-\frac{1}{2}\left\{\frac{[y_k - f(x_k)]^2}{\left[dy_k^2 + f'(x_k)^2 dx_k^2\right]}\right\}\right) \quad (3)$$

where N represents the total number of empirical data points, and the subscript k denotes the kth data point.

For the simple power-law bivariate model $y = ax^b$, the likelihood formula for estimating unknown parameters a and b, using the aforementioned likelihood equation [Eq. (3)], shall be written as follows:

$$\mathrm{Likelihood}[f(x)] = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi\left[dy_k^2 + \left(abx_k^{b-1}dx_k\right)^2\right]}} \exp\left(-\frac{1}{2}\left\{\frac{\left(y_k - ax_k^b\right)^2}{\left[dy_k^2 + \left(abx_k^{b-1}dx_k\right)^2\right]}\right\}\right) \quad (4)$$

In this work, Eq. (4) was used for fitting allometric relationship between body mass and length of Sciuridae species. Because there is no clear consensus on whether logarithmic transformation should be applied for empirical data when fitting bivariate models in either species-area or allometric relationships (Ballantyne 2013; Chen 2009; Connor and McCoy 1979; Manaster and Manaster 1975; Packard 2013), the maximum likelihood model with data uncertainty described herein will be applied on the original data without logarithmic transformation.

The likelihood model [Eq. (3)] can be applied to the situations when any of the variables do not contain measurement uncertainty. If the explanatory variable X is deterministic (i.e. no dx or $dx = 0$, but $dy \neq 0$), the likelihood model [Eq. (3)] will be reformulated as follows:

$$\mathrm{Likelihood}[f(x)] = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi dy_k^2}} \exp\left(-\frac{1}{2}\left\{\frac{[y_k - f(x_k)]^2}{dy_k^2}\right\}\right) \quad (5)$$

By contrast, when the response variable Y is measured without uncertainty (i.e. no dy or $dy = 0$, but $dx \neq 0$), then the likelihood model is written as

$$\mathrm{Likelihood}[f(x)] = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi\left[f'(x_k)^2 dx_k^2\right]}} \exp\left(-\frac{1}{2}\left\{\frac{[y_k - f(x_k)]^2}{\left[f'(x_k)^2 dx_k^2\right]}\right\}\right) \quad (6)$$

Finally, when both X and Y do not contain uncertainty, the likelihood model is reduced to the simple least-of-square model as

$$\mathrm{Likelihood}[f(x)] = \prod_{k=1}^{N} \exp\left(-\frac{1}{2}[y_k - f(x_k)]^2\right) \quad (7)$$

For comparison, the conventional nonlinear least-of-square (NLOS) method is applied on the empirical data sets, but the measurement error terms (dx, dy) are simply ignored when carrying out the fitting procedure on the raw data points (x, y). Consequently, the NLOS method implemented here is equivalent to Eq. (7) described earlier. For the estimation of the parameters, 95% confidence intervals are derived using the Fisher information matrix, which is computed as

$$[I(\theta)] = -E\left(\frac{\partial^2 \log\{\mathrm{Likelihood}[f(x)]\}}{\partial\theta\partial\theta^T}\right) \quad (8)$$

where the parameter vector is represented by $\theta = \{a, b\}$.

Therefore, the asymptotic variance–covariance matrix is constructed as

$$\mathrm{VCV}(\theta) = [I(\theta)]^{-1} \quad (9)$$

where VCV is the variance–covariance matrix.

Thus, the variance for parameters a and b is given by the diagonal elements of the matrix $\mathrm{VCV}(\theta)$, which can be used to compute the 95% confidence interval of the parameters. An *R* script for implementing the aforementioned models is available from the author upon request. The 95% confidence interval for the fitted parameters can also be calculated using simple nonparametric bootstrapping, but it is computer intensive and time consuming.

An empirical data set

Data on body mass (g) and length (mm) for 170 species from the family Sciuridae were derived from a previous study (Hayssen 2008). In the original data set, the body mass and length for each species were measured for female, male, and adult, respectively. However, there were some missing data. Thus, for obtaining standard deviation of the data, species with too many missing data were dropped from the original data set, which eventually resulted in a data matrix with 170 species.

Because each species will have its separate body length and mass data for female, male, and adult, respectively, measure uncertainties are encountered if one wants to infer a general allometric relationship between body mass and length for Sciuridae species regardless of sex and growth status of the species. The general allometric relationship might have to be deduced by applying one of these two methods: (1) by taking the averages of three allometric sex- or growth-biased relationships estimated separately by the NLOS method; or (2) by obtaining a general allometric relationship in order to fit the NLOS method into the mean body mass and length data assembled from the raw data. However, both handling methods will definitely result in loss of information relevant to the body mass and length of species. In the author's comparative study, the second method was adopted.