Journal of Experimental Marine Biology and Ecology 474 (2016) 81–91

Contents lists available at ScienceDirect

Journal of Experimental Marine Biology and Ecology

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

Strengthening statistical usage in marine ecology: Linear regression

Inna Boldina, Peter G. Beninger *

Laboratoire de Biologie Marine, Faculté des Sciences, Université de Nantes, 2, rue de la Houssinière, 44322 Nantes France

ARTICLE INFO

Article history:
Received 3 July 2015
Received in revised form 21 September 2015
Accepted 23 September 2015
Available online 19 October 2015

Keywords:
Regression
Linear
Ecology
Statistics
Assumptions

ABSTRACT

Linear regression is a frequently-used statistical technique in marine ecology, either to model simple relationships or as a component of more complex models. The apparent simplicity of this technique often obscures its far more complex underpinnings, upon which its validity, and ultimate ecological interpretations, wholly depend. We present a non-technical review of the foundations of linear regression and its application in marine ecology, with emphasis on correct model specification, the different concepts of linearity, the issues surrounding data transformation, the assumptions which must be respected, and validation of the regression model. The necessity of reporting the results of regression diagnostics is stressed; contrary to widespread practice in marine ecology, R^2 and p-values alone do not provide sufficient evidence to form conclusions.

Contents

1. Introduction

When analyzing marine ecological data, what could be simpler than a linear regression? Until recently, Excel® would do it without even

* Corresponding author.
E-mail address: Peter.Beninger@univ-nantes.fr (P.G. Beninger).

http://dx.doi.org/10.1016/j.jembe.2015.09.010
0022-0981/

using the term itself ('trend' was so much more user-friendly!). In this ubiquitous statistical technique, as in all others, the devil is not only in the details, but also in the assumptions; for what we have here is a mathematical technique which will always work perfectly in the abstract world of mathematics, but which will never work perfectly, and often will not work very well at all, in the real world. Linear regression is an attempt to describe complex, incompletely understood real-life processes in the simplest and most accurate (aka mathematical) terms possible; at times the correspondence is rather good, but at others, it is like fitting a Phillips screwdriver into a Robertson screw. In other words, the mathematical construct is a model which we hope to use to describe a real-life relation. And in the words of the patriarch of modelization, George Box, 'All models are wrong; some are useful' (Box, 1976).

In a previous paper we attempted to provide guidelines for strengthening statistical usage in marine biology, with the central concerns of frequentist (hypothesis-testing) and inferential approaches (Beninger et al., 2012). In the present work, we wish to address another foundational aspect of statistical analysis in marine ecology: linear regression.

Like all statistical techniques, linear regression is often considered by non-statisticians to be a simple, mechanical tool, performed at the touch of a computer key, without proper consideration of its restrictions, assumptions, and weaknesses, thereby covertly combining ease of operation with ease of error. The purpose of this review is to give a non-technical overview of linear regression principles, as well as the precautions to avoid the most common and serious pitfalls. We pay special attention to the most frequently-violated assumptions of linear regression, in the hope that incorrect usage might diminish in the near future.

At the outset, we must state what linear regression is, and what we hope to accomplish with it, before delving into whether or not we can actually do it, and how.

1.1. What is linear regression?

Regression analysis is a generic term for a group of different statistical techniques. The purpose of all these techniques is to examine the relationship between variables. The most common type of linear regression is Type I regression, in which we attempt to determine the relationship between dependent and explanatory or independent variables. Less well-known is Type II regression, in which there are no independent variables, and all variables can influence each other. A short glossary of the linear regression types is provided in Table 1, and these topics will be developed in the following sections. We will focus on Type I linear regression, which is widely used in many different contexts in aquatic ecology, e.g. the species-area relationship (Begon et al., 1996; Peake and Quinn, 1993), the relationship between population density and body size of benthic invertebrate species (Schmid, 2000), the characterization of spatial patterning (Beninger and Boldina, 2014; Seuront, 2010), the multiple fields in which allometric relations are prominent, e.g. suspension-feeding, population dynamics, metabolic scaling (Carey et al., 2013; Cranford et al., 2011; Gosling, 2015; Hirst, 2012; Robinson et al., 2010), DEB modeling (Duarte et al., 2012; Rosland et al., 2009), relation of phytoplankton cell size and abiotic factors (Finkel et al., 2010), etc. Although this technique is most frequently used to model relationships which are graphically characterized by a straight line, it is important to note that it may also be used to model certain curvilinear relationships (Montgomery and Peck, 1992). This aspect will be explained in Section 2.1.

Table 1
A short glossary of frequently-misunderstood linear regression terms.

Linear regression	Requires linear models (linear in parameters) which may have curvilinear form
Non-linear regression	Requires non-linear models (non-linear in parameters)
Multiple linear regression	Regression with several independent variables
Polynomial linear regression	A special case of multiple linear regression describing a curvilinear relationship
Type I linear regression	Assumes an asymmetrical relationship between dependent and independent variables
Type II linear regression	Assumes a symmetrical relationship between variables; there is no independent variable
Estimator	Function used to calculate the regression equation from the observed data

1.2. What can we accomplish with linear regression?

There are three possible objectives for linear regression analysis in marine ecology:

1) Stating the nature of the relationship between two variables. If our only purpose is to state that 'this is the equation which appears to characterize the relationship', then we have very few preconditions and assumptions to worry about. However, this is not a very useful tool in marine ecology, where we usually wish to predict the value of the dependent variable for a given value of the independent variable (e.g. what sardine or tuna weight corresponds to what sardine or tuna length-values much quicker and easier to measure shipboard?)
2) Dependent variable prediction within the range of observed dependent variables. Here we simply wish to predict any y-value within those corresponding to the maximum and minimum observed x-values, e.g. what weight for any length which falls within the x-coordinates of the maximum and minimum weight values. This is a much more useful objective, but the trade-off is that it requires more, and stricter, assumptions.
3) Dependent variable prediction beyond the range of observed dependent variables. Here we attempt to boldly go where none of our data has gone before, i.e. beyond the maximum and minimum observed y-values. This extension of modeling has been used for everything from enzyme kinetics to climate change. It is usually an attempt to predict a future y-value, something humans have tried to do since they became aware that there is a future. Naturally, this type of objective carries the greatest load of restrictions, assumptions, caveats, and risk of error.

1.3. From the mathematical to the statistical

Linear regression uses the model of a straight line, whose mathematical equation is the familiar.

$$Y = a + bx$$

where a is the y-intercept and b is the slope of the line. Statisticians prefer the notation.

$$Y = \beta_0 + \beta_1 X_1$$

for the population model (Greek letters used by convention), which highlights the fact that the slope and y-intercept are both parameters of the equation.

Much of the very real misunderstanding and misuse of linear regression stems from the widespread tendency of marine ecologists to assume that the abstract, perfect mathematical world can be used to directly model the much messier real world. In the real world, an unknown number of uncontrolled variables other than the independent variable can influence the dependent variable, e.g. individual variations in physiology, handling time of individual samples, or even atmospheric pressure variations. We therefore know that other variables can influence the dependant variable, but we cannot identify them or measure their magnitude. Furthermore, these variables may influence the dependant variables in either an additive fashion (i.e. add their unknown positive or negative values to the linear equation) or in a multiplicative