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

fashion. Statisticians have grouped all the unknown variable contributions under a single term, ε, and given it the unfortunate name of 'error term'. This name is used by convention and for convenience only, a perfect example of mis-naming in the sciences, since much or even most of this variation may not be due to any real 'error', but rather to stochastic events, the differences between individual organisms, etc. (Fox, 2015).

We may therefore modify the linear equation model to better represent the real world thus:

$$Y = \beta_0 + \beta_1 X_{1+} \varepsilon$$

where the error term is assumed to be additive. It will be seen that, far from being an afterthought, the characteristics of the error term are a central issue in regression analysis.

As the foundation for the whole edifice of linear regression, it is worthwhile to examine each term of the regression equation in turn:

- Y is the dependent or outcome variable. It is also called the predicted value, or, most commonly (and inappropriately), the fitted value
- X_1 is the independent or explanatory variable; since it assists in determining the value of the dependent variable, it is also called a predictor
- β_0 and β_1 are the parameters of the regression model:
  - o β_0 is the y-intercept, i.e. where $x = 0$. In real life situations, x rarely equals 0 (there is no biological sense for $x = 0$), so the intercept β_0 has no concrete meaning and serves only to correctly position the regression line with respect to the meaningful data points.
  - o β_1 is the regression coefficient (aka slope of the regression line). The regression coefficient β_1 represents the mean change in the dependent variable for one unit of change of the independent variable.

- ε — the error term — this term is never seen in the equations given in most marine ecology papers, since it cannot be quantified (and its influence and the effects of its characteristics are often unknown or unsuspected), but its properties may hugely influence the equation. It represents all the variation in Y that cannot be explained by the variation in X. This is the mathematical way of quantifying the other real-world effects on the dependent variable (Hoffmann and Shafer, 2015). Each y-value will be affected by this variation, so we can consider an 'error' as the difference between the individual y value in a sample and the unknown true value in the population. Since the effects of ε may range from very small (e.g. in enzyme kinetics) to stupendously large (e.g. in stock-recruitment studies), and vary between y-values from very little to very much, preoccupation with the magnitude, and variation of ε is an extremely important aspect of regression analysis. Since we cannot know the magnitude of ε (by definition unknown), the only available strategy in this case is to reduce it as much as possible, through controls and replication. Similarly, it is usually impossible to characterize the exact variation of ε (aka the error distribution) in a population; however, we can test assumptions about it in our samples, and this will allow us to judge whether or not linear regression is a meaningful way of relating the dependent and independent variables. We do this using the differences between individual y-values in a sample and the corresponding y-values predicted from the regression model (aka residuals, see below and Fig. 3 — Quinn and Keough (2002). Residuals are estimates of the true population error, and the term is often used interchangeably with ε, in the same way that sample standard deviation is used to reflect the population standard deviation. Ideally, residuals should be randomly distributed; in other words, they should not show bias, since statistical techniques are inoperable under conditions of bias. The analysis of residuals is a very important, yet often — neglected part of linear regression. Much more will be said of residuals analysis when we describe correct linear regression procedure.

1.4. Overview of linear regression procedure

Regardless of how quickly and easily a software program will perform a linear regression, the approach comprises three consecutive steps (Fig. 1), with potential pitfalls to be avoided at each step. The first, and the most critical, step is to correctly specify the regression model. The second is to determine the best estimation of the regression parameters, and the third is to validate the model (e.g. test whether the regression parameters are statistically different from zero and verify the goodness of fit). The assumptions concerning the first and second steps can be verified both before and after constructing the regression model, since they concern the form of the model and the independent variables included in it, as well as the distribution of residuals. The third step is post-hoc: validation of the model. In the following section,

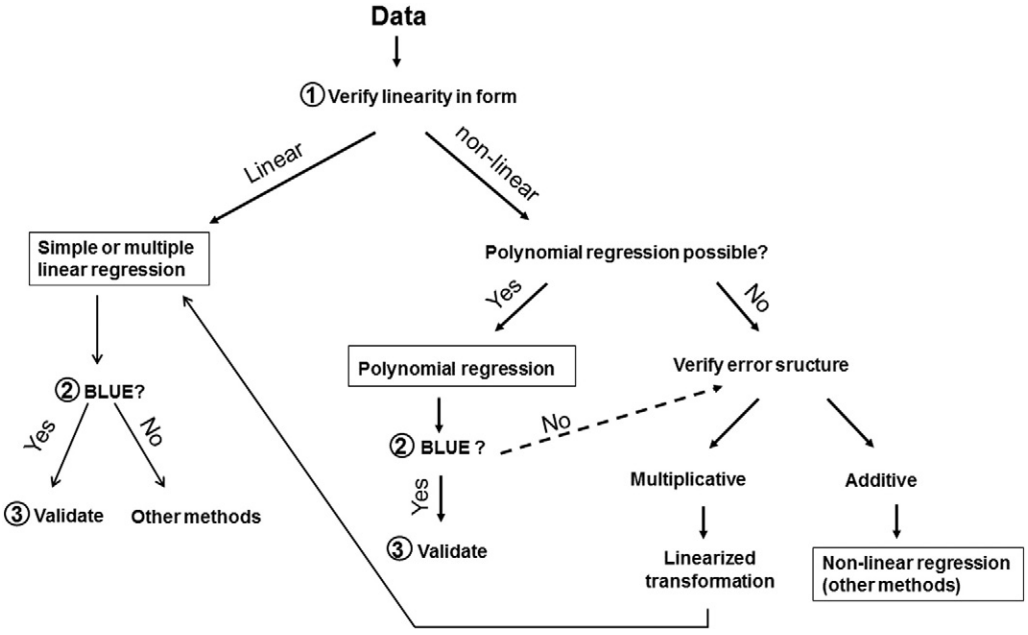

Fig. 1. Summary of linear regression procedure. For explanation of BLUE, see Sections 3.1–3.2.

the assumptions and post-hoc verification for simple and multiple linear regression will be described together, as they are quite similar.

2. Choosing the correct type of regression

Although in the present work we are only interested in linear regression, the following discussion will show that this includes types of regression which analyze relationships other than those characterized by single independent variables and straight lines, and therefore the subject is rather more complex than often perceived. The process is essentially iterative, and includes (1) determining the appropriate functional form (itself dependent on the various 'types' of linearity), and (2) for multiple regression, choosing the explanatory variables to be included. Incorrect procedure may lead to violations of the assumptions and biased parameter estimates (i.e. values which deviate systematically from those of the real world).

2.1. The many faces of linearity

Like all scientific domains, statistics is afflicted with imprecise and often misleading terminology. With respect to linear regression, the very term linear can mean several completely different things, each equally important: linearity in form, linearity in variables, and linearity in parameters. Linearity in form refers to the common, and correct, perception of a linear equation as a straight line. Linearity in variables stipulates that the independent variable(s) (1) do not contain other functions (exponents, trigonometric functions, etc.) and (2) in the case of more than one independent variable, are related to each other by summation (+), e.g. in a multiple linear regression equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Similarly, linearity in parameters stipulates that the regression parameters also do not contain other functions, and are also related linearly to each other (simple summation), as is also the case in the above example. An example of an equation linear in parameters but non-linear in variables would be:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^{\,2} + \varepsilon \qquad \text{(Example 1)}$$

Although one of the terms in this equation is a non-linear variable (X_1^2, a polynomial term), it is possible to use a type of linear regression procedure called polynomial regression. Statistically speaking, polynomial regression is still linear regression, so the underlying assumptions are the same. Polynomial regressions (and particularly quadratic and cubic polynomial) are often used in marine ecology to model parabolic or maximum-minimum relations, respectively (Legendre and Legendre, 2012; Zuur et al., 2007). However, we should note that while it is easy to fit a polynomial model to data, it is often difficult to ascribe any real meaning or rationale to some of the terms, especially for higher-order polynomials, and this is clearly undesirable in any model. A non-linear model is preferable in such cases (Austin, 2007; Müller et al., 2010).

In contrast to non-linearity in variables, an equation may be non-linear in parameters, but nonetheless linear in variables:

$$Y = \beta_0 + \beta_1 \beta_2 X_1 + \beta_2 X_2 + \varepsilon \qquad \text{(Example 2)}$$

Here the regression coefficients are multiplicative and therefore non-linear. Note that, as indicated above, the relationships in these two examples may be (first example) or are (second example) graphically curvilinear. Although in the first example, multiple linear regression may be performed directly, in the second example, other regression techniques are necessary: either linearization via transformation, or non-linear regression (Section 2.4).

In practice, researchers (knowingly or not) must choose the type of regression model prior to performing the regression. If they believe that the relation is linear, they will then perform a linear regression. However, since the relation may simply appear to be linear in form, for example from a scatterplot of raw data, it is absolutely essential to verify this form (see below). The other two "types" of linearity (in variables and parameters) are not verified in practice, but rather fixed by the researcher prior to performing the regression.

2.2. Verification of linearity in form

Contrary to widespread belief, linear regression is not a way of verifying whether a relationship is linear in form. Quite the opposite, it is assumed at the outset that the relationship between dependent and independent variables is linear, and only under this condition can simple linear regression analysis correctly portray the characteristics of this relationship. In practice, this means that we cannot just perform a simple linear regression analysis to see if we obtain a statistically-significant relationship, from which we conclude that the relationship is linear. If simple linear regression is applied to data which present a non-linear relationship, the results will under-estimate the true relationship, producing a biased slope and intercept, thus increasing the degree of error for all three of the objectives of linear regression (Quinn and Keough, 2002).

The simplest solution for verifying linearity in form is to plot the raw data and visually inspect it before choosing the linear model. The scatterplots of some data sets may be too ambiguous for this approach (Fig. 2A); in these cases, a helpful technique is visual examination of the residuals plot, i.e. residuals vs. predicted values (Fig. 2A and B). Note that the horizontal axis in such a plot consists of the predicted (or 'fitted') values (Y).

On the raw data graph, the relationship appears to be approximately linear in form (Fig. 2A). However, if the relationship were truly linear in form, the residuals would be symmetrically distributed around the horizontal line on the residual plot. The considerable departure of linearity at the extremities of the graph indicates that the model makes systematic errors for small and large predictions (Fig. 2B). The superimposed LOESS curve (Locally weighted scatterplot smoothing: a non-parametric technique that graphically represents a curve of best fit without assumptions of data distribution (Cleveland and Devlin, 1988) is not necessary but can facilitate the detection of linearity violation (technical note: an Excel® linear regression add-in allows the construction of residuals vs fitted plots, without the LOESS function, which can be found in R). Where the relation is not linear in form, it is not possible to perform a simple linear regression; one of several other options would be more appropriate, as described below.

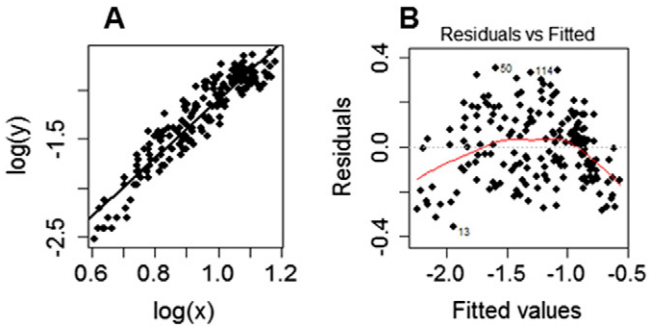

Fig. 2. *Cerastoderma edule*. Unpublished data. A. Log length (x) vs. log weight (y). B. Diagnostic plot of residuals vs fitted values. If the data fit a linear model, the residuals should be grouped near the zero horizontal line. The curved line is a calculated LOESS curve which shows the directions of the distortions from linearity.

2.3. Dealing with non-linearity in form

If the residual plot shows that the relationship is not linear in form, there are several possible avenues for mitigating this problem. The most straightforward are shown in Fig. 1:

1. Polynomial regression on original data (see Section 2.1)
2. Linear regression after linearization of the data through transformation
3. Non-linear regression on original data (a different set of techniques, not considered here)

2.4. Transformation

Since the most common transformation in marine ecology is the log-transformation, especially in allometric studies (Cranford et al., 2011; Ibarrola et al., 2012; Jennings et al., 2001; Katsanevakis et al., 2007; Robinson et al., 2010), it warrants particular attention. If the data plot suggests an exponential rather than a linear relationship between the x and y variables, the corresponding model has the general form:

$$Y = aX^b$$

At this stage of the discussion, we have left out the error term, since we do not yet know whether it is additive or multiplicative. These properties are part of what is known as the error structure.

The choice to be made is whether to log-transform the original data and do linear regression afterwards, or to perform a non-linear regression on the original data. This issue has been vigorously debated with respect to ecological data (Ballantyne, 2013; Glazier, 2013; Kerkhoff and Enquist, 2009; Packard, 2009, 2013). Neither option is demonstrably superior in all cases; rather, the performance of each model depends essentially on the nature of the error term (normal and additive or lognormal and multiplicative — Cohen, 2003; Galton, 1879; Gingerich, 2000; Xiao et al., 2011).

In the case of a model which is non-linear in form, and for which it is not possible to perform a polynomial regression because it is also nonlinear in parameters (Section 2.1), there are two possibilities for the relationship of the error term to the rest of the equation:

1. Additive error term: $Y_i = aX_i^b + \varepsilon_i$
   Where Y_i is the ith observation of the dependent variable, X_i is the ith observation of the independent variable, ε_i the error term associated with the ith data point. The additive error model assumes that the error term has a normal distribution across the range of the independent variable. If a log-transformation is applied to this model (usually when the data are log-transformed without taking the properties of the error term into account), the consequent back-transformation (from logarithmic to real-number values) will be erroneous, thus eliminating the possibility of using a linear regression procedure — including the ubiquitous Ordinary Least Squares (OLS) method, which is explained below (Myers, 1990; Packard and Boardman, 2008). This type of model is called intrinsically nonlinear, because it cannot be linearized (Cohen, 2003), and nonlinear regression on original data is therefore the most appropriate procedure (Gingerich, 2000; Xiao et al., 2011). This is in stark contrast to the widespread practice of log-transformation of data sets characterized by exponential relationships, and in particular the huge literature based on, or involving, allometric data (Xiao et al., 2011).
2. Multiplicative error term
   Linear regression methods on log-transformed data require a multiplicative distribution of residuals of the original data (Kerkhoff and Enquist, 2009; Rawlings et al., 1998):

   $$Y_i = aX_i^b + e^{\varepsilon i}$$

   Where Y_i is the ith observation of the dependent variable, X_i is the ith observation of the independent variable, ε_i the error term associated with the ith data point, and e the natural log base.
   The multiplicative error model assumes that the error term has a lognormal distribution and, in ecological data, its variance usually and naturally increases with an increase in the Y values. In such a model, the error is multiplied by the value of the dependent variable. After taking the log of both sides of the equation we obtain:

   $$\text{Log } Y_i = \log a + b \log X_i + \varepsilon_i$$

   In contrast to the previous situation where the error term was additive, this linearized form of the initial equation can be modeled with linear regression; in other words, this type of model is intrinsically linear. It should be emphasized that log-transformation is an appropriate linearization procedure if and only if the error term is multiplicative. Therefore, whenever log transformation is performed, it should be justified by a clear statement of the nature of the error term; this necessary information is almost universally lacking in the marine ecological literature.
   The next logical step is thus to ascertain whether the error term is additive or multiplicative. However, there are many things we do not, and cannot, directly know about ε, including whether it is additive or multiplicative! We find ourselves in much the same situation as that of particle physicists, who cannot directly visualize sub-atomic particles, but must deduce their existence based on their properties alone. An indirect method has been proposed, which performs both types of regression (non-linear on the raw data and linear on log-transformed data), followed by residuals analysis in both cases (Cohen, 2003; Draper and Smith, 1998). Hence, if the error term is additive and we perform linear regression on log-transformed data, the residuals graph will reveal that one or more assumptions are not satisfied. However, mis-specification of the nature of the error term is only one possible cause of a failed residuals analysis; therefore, a more reliable solution for the determination of error structure via analysis of residuals has been devised, based on the now-familiar AIC criterion, rather than on visual analysis of the residuals (Xiao et al., 2011).
   Despite the necessity of error structure (i.e. residuals) analysis in linear regression, it is very frequently omitted (Xiao et al., 2011), especially in the marine ecology literature, where log-transformation of data is routinely performed without consideration of the error structure. Informal discussions with colleagues at meetings suggest that this fundamental issue is, quite simply, not well known to many marine ecologists.

2.5. Choosing variables to include/multiple linear regression

Since we may be able to identify and test more than one independent variable which influences the dependant variable, we can include them in the equation as

$$Y = \beta_0 + \beta_1 X_{1+} \beta_2 X_{2+} \varepsilon.$$

This type of model is called multiple linear regression. Contrary to simple linear regression, its graphical representation is not straightforward (we can plot partial regression equations, holding several independent variables constant, but no single plot will be feasible beyond three independent variables). We include it in this discussion because technically it is a linear model, and also because the most common linear regression technique, Ordinary Least Squares (OLS) is often used to specify its parameters (Sheather, 2009). Although the model itself can be linear or curvilinear in form, it should not be confused with polynomial regression (Section 2.1).

The determination of independent or explanatory variables which should be included in the multiple linear regression model is governed

by one or more theoretical considerations, prior knowledge, or informed judgment (aka common sense). There are two main sources of error in model specification: omitting relevant variables (underfitting) or including irrelevant variables (overfitting). Obviously, the consequences of underfitting are the most serious, engendering biased regression coefficients — in other words, the resulting regression model is wrong. Overfitting does not bias the regression coefficients, so the regression model remains correct, albeit with a loss of efficiency, and an increase in the probability of a Type II error (Hoffmann and Shafer, 2015). One of the corollaries of George Box's famous 'all models are wrong' statement is

> 'Since all models are wrong, the scientist cannot obtain a 'correct' one by excessive elaboration. On the contrary, following William of Occam, he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist, so overelaboration and overparametrization is often the mark of mediocrity.' (Box, 1976).

3. Estimation of the parameters in the regression model

The next step of the regression analysis is the estimation of the model parameters, based on the collected data. The rule or function used to calculate the regression equation from the observed data is called the estimator. Although there are several possible estimators, each with its merits, the overwhelmingly popular OLS estimator is such a conventional method that in marine ecology, it is virtually synonymous with linear regression itself. The popularity of the OLS method is due partly to particularly convenient characteristics of the OLS estimator (see below), partly to historical reasons (it is one of the oldest statistical methods), and partly to the fact that this technique is the default, and sometimes unique, program in all statistical software packages, including Excel®. It is based on the principle of minimizing the sum of the squared residuals (i.e. squared vertical distances between observed and predicted values — Fig. 3). This technique assumes asymmetry in the nature of the dependent and independent variables (Type I regression), i.e. a change in the independent variable engenders a change in the dependent variable, and not vice versa (Weisberg, 2005), as opposed to Type II regression, where the relationship of the variables is symmetrical (a change in either induces a change in the other — Laws and Archie, 1981; McArdle, 2003). There is a considerable literature concerned with the correct contexts for Type I and Type II regression, especially in the field of allometry (Legendre and Legendre, 2012; Warton et al., 2006).

The OLS method requires that independent variables be either fixed by the study design or be measured without error.

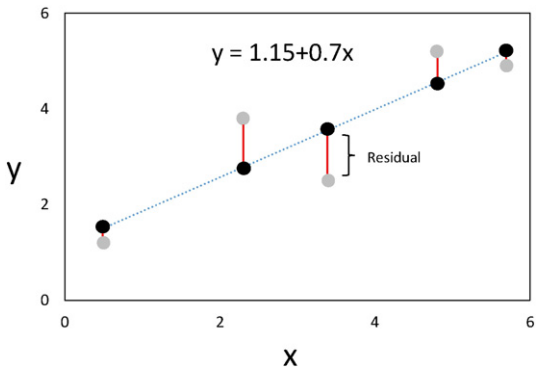

Fig. 3. Modeling of the regression line using the OLS estimator. Gray points: actual values of the dependent variable. Black points: predicted values of the dependent variable.

3.1. Conditions for proper use of OLS

The OLS estimator has several convenient and desirable properties. It is considered 'BLUE' i.e. the Best Linear Unbiased Estimator (Monahan, 2008; Weisberg, 2005) under certain assumptions (Gauss–Markov assumptions, see Section 3.2). To explain these characteristics, OLS is an Estimator because, like many statistical techniques, it is a procedure which estimates the true population parameters of the function, from observed values within random samples of the population (Panik, 2005). It is a Linear estimator because it is a linear function of the dependent Y variables (Bingham and Fry, 2010). It is the Best (i.e. most efficient) estimator of the linear parameters because it produces the least sampling variability (Fig. 4). It is Unbiased because it does not systematically over- or under-estimate the true population parameter β (Fig. 4). In other words, the OLS estimator uses the sample data in the most efficient way for determining the regression line.

The Gauss–Markov theorem states that the OLS estimator can only be BLUE if certain assumptions are met (see below). We must therefore verify all the assumptions, one by one, after performing the regression (aka regression diagnostics). Violation of one or more assumptions means that a better-fitting model could be found. As is always the case, working within the parametric framework means a trade-off between statistical power and severity of constraints. Although it can be very tempting to use OLS without testing the assumptions, since the result is often an estimated regression coefficient (slope) with a low p-value, and hence a seemingly interesting conclusion, the results would in fact be invalid.

Despite the absolute necessity of regression diagnostics, this procedure is either rarely used, or, in the best-case scenario, simply unreported in the marine ecological literature, calling into question the validity of a very great number of published results. Only the systematic testing and reporting of regression diagnostics will improve the reliability of research in which OLS is so frequently used (Faraway, 2014; Quinn and Keough, 2002; Rawlings et al., 1998).

3.2. Striving for BLUE: Gauss–Markov assumptions required for use of OLS

Ideally, our regression estimator should be perfect, i.e. BLUE. OLS is BLUE only if it fulfills a series of conditions known as the Gauss–Markov assumptions (Bingham and Fry, 2010; Demidenko, 2013); this constraint also applies to polynomial regression (which, contrary to common perception, is actually a form of linear regression). Herein we present the assumptions most likely to be violated in marine ecological work.

It is obvious to every biologist that the living world does not completely conform to the ideal world of mathematics (or rather, that we will probably never be able to completely characterize the living world mathematically). It is thus not surprising that ecological data very often fails to satisfy one or more of the assumptions necessary for

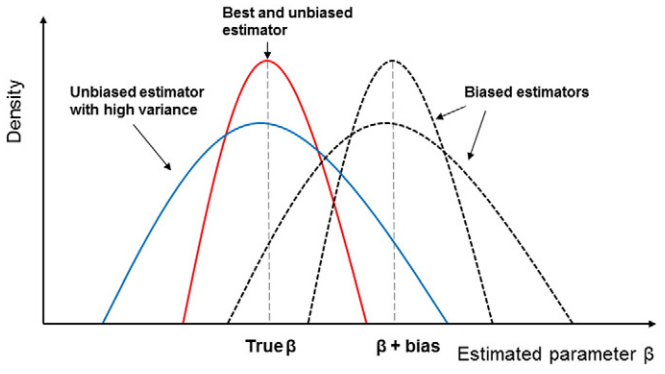

Fig. 4. Sampling distribution of different estimators of the population parameter β.

linear regression of any type, including OLS (Zuur et al., 2009). Nonetheless, it is very important to verify each of these assumptions, and then to understand what happens if one or more assumptions fail, as outlined in the sections that follow. We also provide indications for mitigation of assumption violations and to what extent the regression model can still be used.

3.2.1. Linearity in parameters

As mentioned above (Section 2.1), linear regression requires linearity in parameters, which means that the coefficients of the regression equation must be linear. We mention this again simply because, formally, it is one of the Gauss-Markov assumptions.

3.2.2. Non-perfect collinearity for multiple regression

Two independent or explanatory variables are perfectly collinear when they are intrinsically related by a linear relationship, and we can compute the value of one variable if we know the value of another. One of the variables is therefore redundant and must be removed from the regression equation (Zuur et al., 2007). Note that while the variables should not be related by linear relationship, they can be related by a non-linear (e.g. quadratic) relationship. In practice, this requires a knowledge of the relationships between variables which we may or may not possess, e.g. salinity and conductivity.

The remaining pertinent Gauss–Markov assumptions are all relevant to the residuals, rather than to the data themselves.

3.2.3. Correlation between the independent variables and the error term

This is one of the most important assumptions concerning the residuals, but probably the least known (Weisberg, 2005). It is formally called the 'Zero conditional mean' assumption, although this unfortunate designation gives no easily-understandable indication of its meaning. In simple terms: all of the independent variables should be uncorrelated with the error term. If the error term changes with the independent variable X (i.e. if there is a relationship between them), the variation in the dependent variable Y is influenced not only by the change in the independent variable X but also by the change in the error term. Such a correlation will result in biased estimates of the regression coefficient.

The assumption of zero conditional mean fails if an important independent variable is omitted or if the functional relationship is mis-specified (the first step of regression analysis). The violation of this assumption most often occurs when the independent variables are not included in the correct algebraic form. An example would be a study of the effect of abiotic factors on the larval development of shrimp where one of the explanatory variables is water temperature. If the true relationship includes the temperature as well as the squared term of the temperature, accounting for the observed non-linear effect of temperature, then omitting to include the squared term of temperature will result in a mis-specified functional relationship.

This crucial assumption may be checked by plotting residuals against the independent variable (X). In an allometric study relating shell length and dry weight of cockles, for example (Fig. 5), it is clear that in the middle and right regions of the graph, the data points are grouped below the zero line, so there is a relationship between residuals and x values, therefore violating the assumption of zero conditional mean.

Such a violation is easily corrected by putting the independent variables into correct algebraic form, and/or including omitted variables (in the case of multiple linear regression).

3.2.4. Absence of autocorrelation in the error term

The presence of autocorrelation in the error term ('officially', and once again ambiguously, called independence of residuals) means that some part of the error term exhibits correlation with another part. Since this is definitely not a random variation, it implies that some useful information has not been included in the regression model (Hoffmann and Shafer, 2015; Quinn and Keough, 2002). This type of

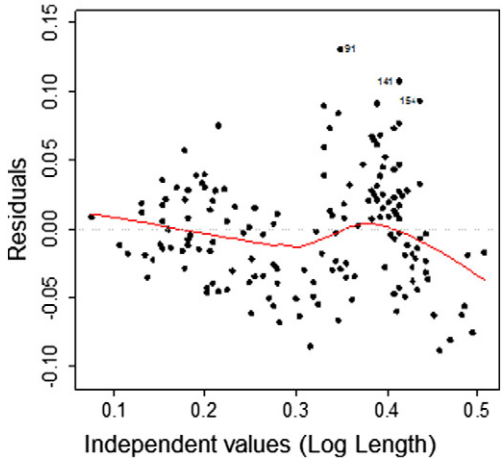

Fig. 5. Diagnostic plot for testing the assumption of zero conditional mean for the relation log weight–log length in the cockle, *Cerastoderma edule*. Unpublished data.

situation is often encountered in marine ecology, notably with data collected repeatedly over time (time series data) and with data collected across spatial units such as transects, quadrats, etc. (Boldina and Beninger, 2013; Fortin and Dale, 2005; Legendre and Legendre, 2012; Whitton et al., 2015). Usually there are stronger associations among errors in adjacent time/space periods than in those that are farther apart (Legendre and Legendre, 2012).

Although autocorrelation in residuals does not affect the unbiasedness of the OLS estimator, it no longer has a minimum variance, which is necessary for BLUE (Fig. 4). In such a situation, the forecasts, confidence intervals, and scientific insights yielded by a regression model may be (at best) inefficient or (at worst) seriously biased or misleading (Glover et al., 2011).

Positive autocorrelation in residuals (the most frequent case in marine ecology) is characterized by standard errors of the estimated parameters which are smaller than the true standard errors. This increases the probability of type I error, i.e. unwarranted rejection of the null hypothesis (Cowpertwait and Metcalfe, 2009; Dale and Fortin, 2002; Legendre, 1993; Zuur et al., 2010).

The assumption of independence of residuals is one of the most commonly - violated assumptions in marine ecology, probably due to a lack of recognition of the importance of testing for the presence of autocorrelation in residuals. The same situation is observed in terrestrial ecology (Beale et al., 2010; Dormann, 2007), traditionally a field with a stronger statistical culture.

The simplest way to detect autocorrelation in residuals is to analyze the plot of residuals against time or sequential order (Fig. 6).

The presence of autocorrelation in residuals (positive, negative or cyclical) is not always obvious from the residuals plot. It is therefore prudent to perform a test for autocorrelation in residuals, of which the most common is the Durbin–Watson statistic (Durbin and Watson, 1951) available in Excel®. The value of the Durbin–Watson statistic ranges from 0 to 4, and the rule of thumb is that the residuals are not correlated if the Durbin–Watson statistic is approximately 2. It should be noted that this test is used only for autocorrelation between adjacent values (aka first-order correlation — Chatterjee and Hadi, 2012). Although this is the most frequent case in marine ecology (e.g. adjacent sampling dates, adjacent sampling sites), second- or third-order autocorrelation, even without first-order autocorrelation, may also be encountered. Other patterns of autocorrelation can be detected using the more sophisticated techniques of Moran'slfor spatial autocorrelation (Legendre, 1993) and autocorrelation function for time series data (Kirchgässner and Wolters, 2007).

If the assumption of absence of autocorrelation in the error term fails, a variety of statistical techniques exist which are designed for

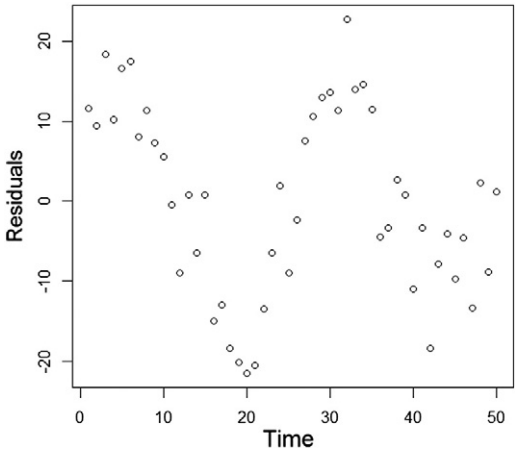

Fig. 6. Example of a residual plot for testing the presence of autocorrelation in residuals. Data simulated in R.

modeling data with autocorrelated errors. Although detailed description of these methods is beyond the scope of this paper, interested readers are referred to techniques such as ARIMA, wavelets or mixed — effect models for time series data (Demidenko, 2013; Kirchgässner and Wolters, 2007), or Generalized Least Squares-S, Simultaneous Autoregressive Models, Generalized Additive Mixed Models and Bayesian Conditional Autoregressive models for spatial data (Beale et al., 2010; Sheather, 2009).

3.2.5. Homoscedasticity

Homoscedasticity (aka homogeneity of variance) means that the variance of errors is the same for all values of the independent variable. This is essentially equivalent to the distribution of residuals, which we may determine once again from a plot of residuals (residuals vs 'fitted' y-values). We emphasize again that the assumption of homoscedasticity for linear models is not about the independent and dependent variables (Y or X), but about the error term, represented by the residuals (Piegorsch and Bailer, 2005).

Regrettably, heteroscedastic errors in linear regression models are a common occurrence in marine ecology. Heteroscedasticity is frequently encountered in data collected at one point in time (aka cross-sectional data). This type of data often contains heteroscedasticity problems for diverse reasons: because marine organisms often have an aggregated, rather than a random, spatial distribution (Beninger and Boldina, 2014; Boldina and Beninger, 2013, 2014; Boldina et al., 2014; Zuur et al., 2007), or because of sexual dimorphism, subpopulation differences, or model mis-specification (omitting important variables from the model or including variables in incorrect algebraic form — Piegorsch and Bailer, 2005; Zuur et al., 2009).

Heteroscedasticity in residuals does not result in biased parameter estimates; however, once again the OLS estimator no longer has a minimum variance, which is necessary for BLUE (Fig. 4). In such a case, another estimator with lower variance can be used, for example the GLS estimator (Demidenko, 2013). The violation of the assumption of homoscedasticity produces unreliable standard error estimates of the parameters. This in turn leads to bias in test statistics and confidence intervals, which increases the probability of a Type I error (Härdle, 2004; Zuur et al., 2009). Although slight heteroscedasticity has minimal effect on significance tests, severe heteroscedasticity is quite problematic (Tabachnick and Fidell, 2007). It is impossible to define exactly when heteroscedasticity becomes a serious issue (this is once again a matter of individual judgment) but in general the linear regression model becomes doubtful when the largest variance is more than four times the lowest variance (Fox, 2008).

The simplest method to check this assumption is, once again, the visual examination of residuals plotted against 'fitted' values (Zuur et al., 2010). If the assumption of homoscedasticity is met, the pattern of residuals will have approximately the same spread on both sides of the horizontal line drawn through the average residual (Fig. 7A). Heteroscedasticity is frequently manifested on the plot by a funnel pattern with non-constant error variance (Fig. 7B).

The presence of heteroscedasticity in residuals is not always visually obvious from the residuals plot alone (e.g. for large data sets), so more formal tests for heteroscedasticity should be performed (Wright and London, 2009). The choice of statistical test depends on the knowledge we have about the characteristics of our data: the Breusch–Pagan test is most suited to the detection of linear forms of heteroscedasticity (Breusch and Pagan, 1979), while White's general test for heteroscedasticity (White, 1980) is more suited to the detection of non-linear forms of heteroscedasticity or when the errors (i.e. residuals) are non-normally distributed (Evans, 1992).

There are several strategies to deal with heteroscedasticity in residuals. The first step is to verify if the model specification is correct (the first step of linear regression). Sometimes including the variable in correct form may resolve the issue. In cases where the independent variable (X) is skewed, transformation is the simplest solution to the problem (Hoffmann and Shafer, 2015). If neither of these approaches succeeds in eliminating the heteroscedasticity, there are a variety of statistical methods which can deal with heteroscedastic residuals, such as Generalized Least Squares (GLS — Zuur et al., 2007), Weighted Least Squares (WLS — Sheather, 2009), and robust standard errors (Vittinghoff, 2011).

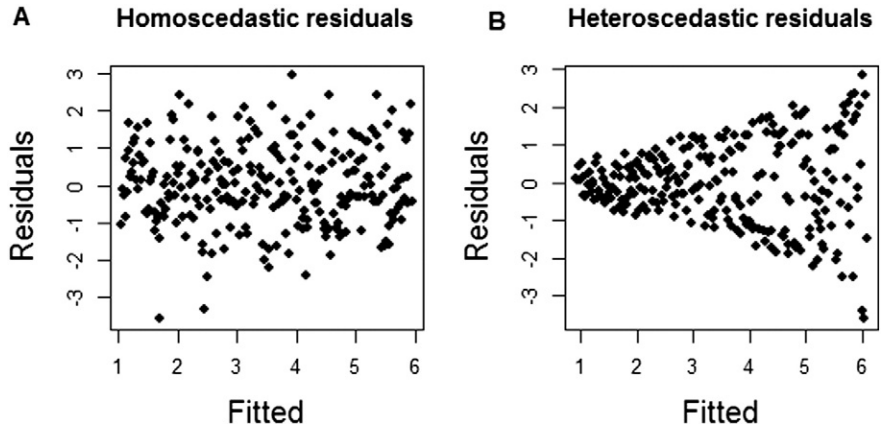

Fig. 7. Residual plot for testing homoscedasticity of residuals. A. Constant distribution (spread) of residuals B. Heteroscedasticity apparent: the spread of residuals increases as the fitted values increase. Data simulated in 'R'.

As mentioned above, marine ecological data is very often afflicted with heteroscedasticity in residuals. Notwithstanding, residual analyses are rarely reported, and it is quite probable that many, if not most, of the published linear regressions performed using the OLS method suffer, to various degrees, from the consequences of heteroscedasticity in residuals. We strongly recommend that marine ecologists routinely check the assumption of heteroscedasticity, and familiarize themselves with alternate regression methods such as GLS when deviation from this assumption cannot be mitigated.

3.3. Assumption of normality

It is a popular misbelief that the dependent variable in linear regression must be normally distributed, perhaps because a non-normally distributed dependent variable often results in violation of homoscedasticity and linearity assumptions (Quinn and Keough, 2002). Nevertheless, in linear regression the assumption of normality only concerns the error term, exactly as for the assumptions of homoscedasticity and non-correlated residuals (Zuur et al., 2007). For the record, we should emphasize that assumption of normality is not a Gauss-Markov assumption. Hence, violation of this assumption is less critical than violation of the Gauss–Markov assumptions. Furthermore, normality of the residuals' distribution is only necessary for the validity of subsequent parametric tests on the regression coefficients, whereas the correct estimation of the coefficients themselves only requires that the errors be neither autocorrelated nor homoscedastic (Zuur et al., 2007). Moreover, the assumption of normality in residuals distribution is critical only in small data sets, which is not often the case in marine ecology. In large samples, hypothesis tests are robust against violations of the normality assumption because of the central limit theorem (Cottingham et al., 2005; Fitzmaurice et al., 2011; Stewart-Oaten, 1995).

Violation of the assumption of normality may have several causes: mis-specification of the model (e.g. when linear regression is applied to a non-linear relationship), the presence of outliers (frequent in ecological studies), or some cases of severe skewness of the dependent variables. Once again, we can visually check the normality assumption using residual plots, but the usual procedure is the Q–Q plot (quantile–quantile plot), because the departure from normality is more apparent, and it is possible to determine which part of the distribution is not fitted correctly. If the residuals are normally distributed, the points in the Q–Q-normal plot lie on a straight diagonal line; obviously, a slight departure from this line is acceptable (Fig. 8A). Assumption of normality fails when the residuals exhibit curvature about the diagonal line (Fig. 8B).

The results of visual plot inspections should be carefully interpreted in the light of the sample sizes. Very small samples give insufficient information on the exact shape of the residuals distribution, whereas in large samples, even small departures from normality may be disproportionately visible in plots (Reimann, 2008).

If there is a need for further statistical analysis requiring normal residuals distribution, transformations of independent variables may normalize the distribution. In the case of large data sets with strong non-normality, generalized linear models (GLM) are recommended (Bolker et al., 2009; McCullagh and Nelder, 1989).

4. Validation of the regression model

Model validation is the last step of the regression analysis, and the one with which most marine ecologists are at least partially familiar. Unfortunately, in marine ecology this crucial step is often limited to exhibiting a high R^2 value. *Considering the R^2 as the unique, or even main, benchmark for model fit is the most common, and most egregious, pitfall in linear regression.*

R^2 (aka coefficient of determination) represents the proportion of the total variance in the dependent variable Y "explained" by linear regression model:

$$R^2 = 1 - \text{SS res/SS total}.$$

As such, R^2 is a statistical measure of goodness of fit of the regression model, i.e. it measures how close the data are to the fitted regression line. R^2 has two serious limitations: it does not reveal whether the regression model is correctly specified (Section 2), or whether the necessary assumptions have been respected (Section 3.2); the possible consequence is the calculation of biased coefficient estimates (Weisberg, 2005). Consequently, a high value of R^2 does not in itself suffice to validate the regression model. This critical point was made most elegantly by Anscombe (1973) (the famous 'Anscombe's quartet'), and it is equally pertinent today. It is thus obvious that reporting an R^2 (even with a p-value), with no information on the regression diagnostics, has no scientific value, and this practice should cease completely.

In multiple and polynomial regression, high R^2 values may occur when the regression model is overfitted. Regression models with more additional independent variables or higher - order polynomials will naturally have higher R^2 values (Sheather, 2009). Consequently, a model with more predictors may appear to have a better fit simply because it has more terms. In reality, overfitted models produce misleadingly high R^2 because they model the random noise component of

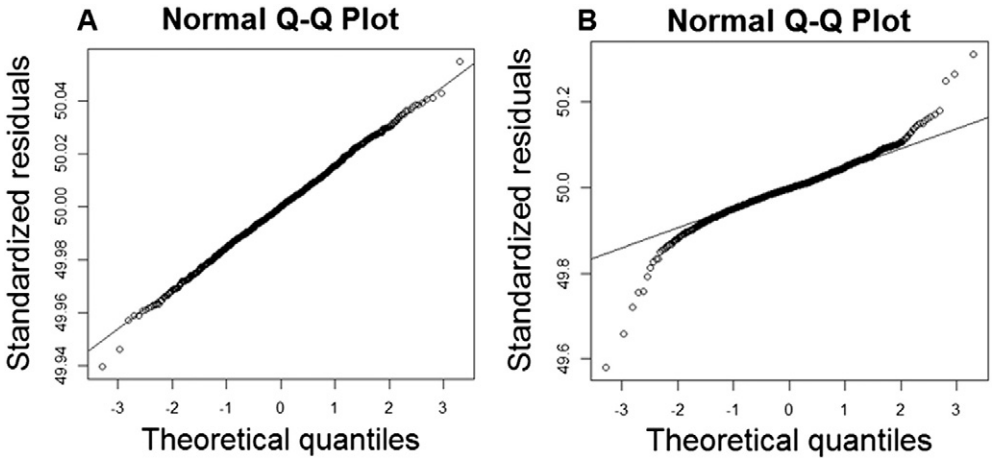

Fig. 8. Q–Q plot for testing normality of residuals. A. Normal distribution of residuals. B. Violation of normality assumption. Data simulated in R.

the error term. Consequently, a much more appropriate statistic for multiple linear and polynomial regression is the Adjusted R^2, which takes into account not only the goodness of fit of the model, but also the number of predictors (Cohen, 2003).

Among the techniques for validating the regression model is the t-test for regression coefficients, based on the t-score:

$$t = b_1/SE$$

where b_1 is the slope (regression coefficient) of the sample regression line, and SE is the standard error of the slope. For multiple and polynomial linear regression, the ANOVA F-ratio test is widely used (Quinn and Keough, 2002). As mentioned above, however, the results of these tests are only valid if all the regression assumptions are satisfied. Furthermore, if the data set is large enough, the obtained t-values may often be statistically significant (Beninger et al., 2012), regardless of the goodness of fit. These validation statistics should therefore be interpreted only in the light of the information concerning the model specification and regression diagnostics; *This information should be supplied, or made available, by the authors of every manuscript in which linear regression is used*.

5. Conclusion

From the foregoing, it should be apparent that the 'simple' technique of linear regression, when correctly performed, is actually a somewhat more complex process, whose successful application requires several verification procedures. The most serious and common errors encountered in the marine ecological literature are summarized in Table 2. Incorrect linear regression may involve both errors in model specification, as well as errors of omission if regression diagnostics are not performed and reported. Residuals analysis is a neglected, but nonetheless key component of this verification process. Similarly, for a linear regression model to be accepted as evidence, it is essential that validation go beyond the conventional reporting of R^2 and p-values, to include the results of the regression diagnostics. The absence of this information from much of the marine ecological literature constitutes a serious source of doubt concerning the results and conclusions of all such past studies; it is hoped that the present paper will help redress this situation for future research.

Table 2
A rogue's gallery of the most serious/common errors in linear regression.

1) The error term of a regression equation is essentially useless, because the published equation never includes it.
2) The characteristics of the error term have no bearing on the regression model or on regression procedure.
3) It does not matter whether the model is actually linear in form, as long as the R^2 value is high.
4) If the data are not normally distributed, a straightforward linear regression is not possible.
5) Residuals analysis is an esoteric procedure not relevant to, or necessary for, everyday linear regression in marine ecology.
6) Data sets involving exponential terms should always be log-transformed prior to regression analysis.
7) The more explanatory variables we include in a multiple linear regression model, the more efficient it becomes.
8) Performing linear regression on time-series data presents no particular problem of validity.
9) Performing linear regression on data collected across spatial units presents no particular problem of validity.
10) Heteroscedasticity only concerns the data set, and in any event is not a common linear regression problem in marine ecology.
11) Reporting R^2 and p-values constitutes a meaningful representation of the validity of a linear regression model.
12) A high R^2 and low p-values are convincing evidence of a valid linear relation.

Acknowledgments

We thank the Région Pays de la Loire for funding during the course of this work (EPAT 2015-02488), as well as one of the reviewers for particularly pertinent comments. **[SS]**

References

Anscombe, F.J., 1973. Graphs in statistical analysis. Am. Stat. 27 (1), 17–21.
Austin, M., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol. Model. 200 (1–2), 1–19.
Ballantyne, F., 2013. Evaluating model fit to determine if logarithmic transformations are necessary in allometry: a comment on the exchange between Packard (2009) and Kerkhoff and Enquist (2009). J. Theor. Biol. 317, 418–421.
Beale, C.M., Lennon, J.J., Yearsley, J.M., Brewer, M.J., Elston, D.A., 2010. Regression analysis of spatial data. Ecol. Lett. 13 (2), 246–264.
Begon, M., Harper, J.L., Townsend, C.R., 1996. Ecology: Individuals, Populations, and Communities. 3rd ed. Blackwell Science, Oxford, Cambridge, Mass. (1068 pp.).
Beninger, P.G., Boldina, I., 2014. Fine-scale spatial distribution of the temperate infaunal bivalve *Tapes* (=*Ruditapes*) *philippinarum* (Adams and Reeve) on fished and unfished intertidal mudflats. J. Exp. Mar. Biol. Ecol. 457, 128–134.
Beninger, P.G., Boldina, I., Katsanevakis, S., 2012. Strengthening statistical usage in marine ecology. J. Exp. Mar. Biol. Ecol. 426-427, 97–108.
Bingham, N.H., Fry, J.M., 2010. Regression: Linear Models in Statistics. Springer, London, New York (284 pp.).
Boldina, I., Beninger, P.G., 2013. Fine-scale spatial structure of the exploited infaunal bivalve Cerastoderma edule on the French Atlantic coast. J. Sea Res. 76, 193–200.
Boldina, I., Beninger, P.G., 2014. Fine-scale spatial distribution of the common lugworm *Arenicola marina*, and effects of intertidal clam fishing. Estuar. Coast. Shelf Sci. 143, 32–40.
Boldina, I., Beninger, P.G., Le Coz, M., 2014. Effect of long-term mechanical perturbation on intertidal soft-bottom meiofaunal community spatial structure. J. Sea Res. 85, 85–91.
Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H., White, J.-S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24 (3), 127–135.
Box, G.E., 1976. Science and statistics. J. Am. Stat. Assoc. 71 (356), 791–799.
Breusch, T.S., Pagan, A.R., 1979. A simple test for heteroscedasticity and random coefficient variation. Econometrica 7 (5), 1287–1294.
Carey, N., Sigwart, J.D., Richards, J.G., 2013. Economies of scaling: more evidence that allometry of metabolism is linked to activity, metabolic rate and habitat. J. Exp. Mar. Biol. Ecol. 439, 7–14.
Chatterjee, S., Hadi, A.S., 2012. Regression Analysis by Example. Wiley, Hoboken, New Jersey (393 pp.).
Cleveland, W.S., Devlin, S.J., 1988. Locally weighted regression: an approach to regression analysis by local fitting. J. Am. Stat. Assoc. 83 (403), 596–610.
Cohen, J., 2003. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. 3rd ed. L. Erlbaum Associates, Mahwah (703 pp.).
Cottingham, K.L., Lennon, J.T., Brown, B.L., 2005. Knowing when to draw the line: designing more informative ecological experiments. Front. Ecol. Environ. 3, 145–152.
Cowpertwait, P.P., Metcalfe, A.V., 2009. Introductory Time Series with R. Springer-Verlag New York Inc., New York (256 pp.).
Cranford, P.J., Ward, J.E., Shumway, S.E., 2011. Bivalve filter feeding: variability and limits of the aquaculture biofilter. In: Shumway, S.E. (Ed.), Shellfish Aquaculture and the Environment. Wiley-Blackwell, pp. 81–124.
Dale, M., Fortin, M.-J., 2002. Spatial autocorrelation and statistical tests in ecology. Ecoscience 9, 162–167.
Demidenko, E., 2013. Mixed Models: Theory and Applications with R. Wiley-Blackwell (754 pp.).
Dormann, C.F., 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Glob. Ecol. Biogeogr. 16 (2), 129–138.
Draper, N.R., Smith, H., 1998. Applied Regression Analysis. 3rd ed. Wiley, New York (706 pp.).
Duarte, P., Fernández-Reiriz, M.J., Labarta, U., 2012. Modelling mussel growth in ecosystems with low suspended matter loads using a Dynamic Energy Budget approach. J. Sea Res. 67 (1), 44–57.
Durbin, J., Watson, G.S., 1951. Testing for serial correlation in least squares regression. II. Biometrika 38 (1/2), 159–177.
Evans, M., 1992. Robustness of size of tests of autocorrelation and heteroscedasticity to nonnormality. J. Econ. 51 (1–2), 7–24.
Faraway, J.J., 2014. Linear Models with R. 2nd ed. Taylor & Francis (286 pp.).
Finkel, Z.V., Beardall, J., Flynn, K.J., Quigg, A., Rees, T.A.V., Raven, J.A., 2010. Phytoplankton in a changing world: cell size and elemental stoichiometry. J. Plankton Res. 32 (1), 119–137.
Fitzmaurice, G.M., Laird, N.M., Ware, J.H., 2011. Applied Longitudinal Analysis. 2nd ed. Wiley, Hoboken, N.J. (701 pp.).
Fortin, M.-J., Dale, M., 2005. Spatial Analysis: A Guide for Ecologists. Cambridge University Press, Cambridge, UK (365 pp.).
Fox, J., 2008. Applied Regression Analysis and Generalized Linear Models. SAGE Publications (688 pp.).
Fox, G.A., 2015. Ecological Statistics: Contemporary Theory and Application. Oxford University Press, Oxford (400 pp.).
Galton, F., 1879. The geometric mean in vital and social statistics. Proc. R. Soc. Lond. 29, 365–367.
Gingerich, P.D., 2000. Arithmetic or geometric normality of biological variation: an empirical test of theory. J. Theor. Biol. 204 (2), 201–221.

Glazier, D.S., 2013. Log-transformation is useful for examining proportional relationships in allometric scaling. J. Theor. Biol. 334, 200–203.
Glover, D.M., Jenkins, W.J., Doney, S.C., 2011. Modeling Methods for Marine Science. Cambridge University Press, Cambridge, New York, N.Y. (571 pp.).
Gosling, E.M., 2015. Marine Bivalve Molluscs. Wiley-Blackwell (536 pp.).
Härdle, W., 2004. Nonparametric and Semiparametric Models. Springer, Berlin, New York (299 pp.).
Hirst, A.G., 2012. Intraspecific scaling of mass to length in pelagic animals: ontogenetic shape change and its implications. Limnol. Oceanogr. 57 (5), 1579–1590.
Hoffmann, J.P., Shafer, K., 2015. Linear Regression Analysis: Assumptions and Applications. NASW Press (226 pp.).
Ibarrola, I., Arambalza, U., Navarro, J.M., Urrutia, M.B., Navarro, E., 2012. Allometric relationships in feeding and digestion in the Chilean mytilids *Mytilus chilensis* (Hupé), *Choromytilus chorus* (Molina) and *Aulacomya ater* (Molina): a comparative study. J. Exp. Mar. Biol. Ecol. 426–427, 18–27.
Jennings, S., Kaiser, M.J., Reynolds, J.D., 2001. Marine Fisheries Ecology. Blackwell Science, Oxford, Malden, MA, USA (417 pp.).
Katsanevakis, S., Thessalou-Legaki, M., Karlou-Riga, C., Lefkaditou, E., Dimitriou, E., Verriopoulos, G., 2007. Information-theory approach to allometric growth of marine organisms. Mar. Biol. 151 (3), 949–959.
Kerkhoff, A.J., Enquist, B.J., 2009. Multiplicative by nature: why logarithmic transformation is necessary in allometry. J. Theor. Biol. 257 (3), 519–521.
Kirchgässner, G., Wolters, J., 2007. Introduction to Modern Time Series Analysis. Springer, Berlin (274 pp.).
Laws, E.A., Archie, J.W., 1981. Appropriate use of regression analysis in marine biology. Mar. Biol. 65 (1), 13–16.
Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74 (6), 1659–1673.
Legendre, P., Legendre, L., 2012. Numerical Ecology. 3rd ed. Elsevier, Amsterdam, Boston (990 pp.).
McArdle, B.H., 2003. Lines, models, and errors: regression in the field. Limnol. Oceanogr. 48, 1363–1366.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. 2nd ed. Chapman and Hall, London, New York (511 pp.).
Monahan, J.F., 2008. A Primer on Linear Models. Chapman & Hall/CRC, Boca Raton (287 pp.).
Montgomery, D.C., Peck, E.A., 1992. Introduction to Linear Regression Analysis. 2nd ed. Wiley, New York (527 pp.).
Müller, F., Baessler, C., Schubert, H., Klotz, S., 2010. Long-term ecological research. Springer (456 pp.).
Myers, R.H., 1990. Classical and Modern Regression with Applications. 2nd ed. Duxbury/Thompson Learning, Pacific Grove, CA. (488 pp.).
Packard, G.C., 2009. On the use of logarithmic transformations in allometric analyses. J. Theor. Biol. 257 (3), 515–518.
Packard, G.C., 2013. Fitting statistical models in bivariate allometry: scaling metabolic rate to body mass in mustelid carnivores. Comp. Biochem. Physiol. A Mol. Integr. Physiol. 166 (1), 70–73.
Packard, G.C., Boardman, T.J., 2008. Model selection and logarithmic transformation in allometric analysis. Physiol. Biochem. Zool. 81 (4), 496–507.
Panik, M.J., 2005. Advanced Statistics from an Elementary Point of View. Academic Press, Oxford (824 pp.).
Peake, A.J., Quinn, G.P., 1993. Temporal variation in species-area curves for invertebrates in clumps of an intertidal mussel. Ecography 16 (3), 269–277.
Piegorsch, W.W., Bailer, A.J., 2005. Analyzing Environmental Data. Wiley, Chichester, West Sussex, England, Hoboken, NJ (496 pp.).
Quinn, G.P., Keough, M.J., 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press, Cambridge, UK, New York (537 pp.).
Rawlings, J.O., Pantula, S.G., Dickey, D.A., 1998. Applied Regression Analysis: A Research Tool. 2nd ed. Springer, New York (657 pp.).
Reimann, C., 2008. Statistical Data Analysis Explained: Applied Environmental Statistics with R. John Wiley & Sons, Chichester, England, Hoboken, NJ (343 pp.).
Robinson, L.A., Greenstreet, S.P.R., Reiss, H., Callaway, R., Craeymeersch, J., de Boois, I., Degraer, S., Ehrich, S., Fraser, H.M., Goffin, A., Kröncke, I., Jorgenson, L.L., Robertson, M.R., Lancaster, J., 2010. Length–weight relationships of 216 North Sea benthic invertebrates and fish. J. Mar. Biol. Assoc. U. K. 90 (01), 95–104.
Rosland, R., Strand, Ø., Alunno-Bruscia, M., Bacher, C., Strohmeier, T., 2009. Applying dynamic energy budget (DEB) theory to simulate growth and bio-energetics of blue mussels under low seston conditions. J. Sea Res. 62 (2–3), 49–61.
Schmid, P.E., 2000. Relation between population density and body size in stream communities. Science 289 (5484), 1557–1560.
Seuront, L., 2010. Fractals and Multifractals in Ecology and Aquatic Science. CRC Press/Taylor & Francis, Boca Raton (344 pp.).
Sheather, S.J., 2009. A Modern Approach to Regression with R. Springer, New York, London (392 pp.).
Stewart-Oaten, A., 1995. Rules and judgments in statistics: three examples. Ecology 76 (6), 2001–2009.
Tabachnick, B.G., Fidell, L.S., 2007. Using Multivariate Statistics. 5th ed. Pearson/Allyn & Bacon, Boston (980 pp.).
Vittinghoff, E., 2011. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. 2nd ed. Springer (509 pp.).
Warton, D.I., Wright, I.J., Falster, D.S., Westoby, M., 2006. Bivariate line-fitting methods for allometry. Biol. Rev. Camb. Philos. Soc. 81 (2), 259–291.
Weisberg, S., 2005. Applied Linear Regression. 3rd ed. Wiley-Interscience, Hoboken, N.J. (310 pp.).
White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48 (4), 817–838.
Whitton, T.A., Jenkins, S.R., Richardson, C.A., Hiddink, J.G., 2015. Changes in small scale spatial structure of cockle *Cerastoderma edule* (L.) post-larvae. J. Exp. Mar. Biol. Ecol. 468, 1–10.
Wright, D.B., London, K., 2009. Modern Regression Techniques Using R: A Practical Guide for Students and Researchers. SAGE, Los Angeles, London (204 pp.).
Xiao, X., White, E.P., Hooten, M.B., Durham, S.L., 2011. On the use of log-transformation vs. nonlinear regression for analyzing biological power laws. Ecology 92 (10), 1887–1894.
Zuur, A.F., Ieno, E.N., Smith, G.M., 2007. Analysing Ecological Data. Springer, New York, London (672 pp.).
Zuur, A., Ieno, E.N., Walker, N., Saveliev, A.A., Smith, G.M., 2009. Mixed Effects Models and Extensions in Ecology with R. Springer-Verlag, New York (574 pp.).
Zuur, A.F., Ieno, E.N., Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems. Methods Ecol. Evol. 1 (1), 3–14.