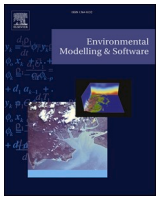
Environmental Modelling and Software 192 (2025) 106526

Contents lists available at ScienceDirect

Environmental Modelling and Software

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



Position paper: Common mistakes and solutions for a better use of correlation- and regression-based approaches in environmental sciences

Damien Tedoldi [a,*], Boram Kim [a], Santiago Sandoval [b], Nicolas Forquet [c], Bruno Tassin [b]

[a] INSA Lyon, DEEP, UR7429, Villeurbanne, 69621, France
[b] LEESU, ENPC, Institut Polytechnique de Paris, Univ Paris Est Créteil, Champs-sur-Marne, France
[c] INRAE, UR REVERSAAL, Villeurbanne, 69625, France

ARTICLE INFO

Keywords:
Bivariate analysis
Data-driven modelling
Empirical modelling
Good practices
Linear regression
Statistical testing

ABSTRACT

While empirical modelling remains a popular practice in environmental sciences, an alarming number of misuses of correlation- and regression-based techniques are encountered in recent research, although these techniques are described in courses and textbooks. This position paper reviews the most common issues, and provides theoretical background for understanding the interests and limitations of these methods, based on their underlying assumptions. We call for a reconsideration of misleading practices, including: the application of linear regression to data points that do not display a linear pattern, the failure to pinpoint influential points, the inappropriate extrapolation of empirical relationships, the overrated search for "statistical significance", the pooling of data belonging to different populations, and, most importantly, calculations without data visualization. We urge reviewers to be vigilant on these aspects. We also recall the existence of alternative approaches to overcome the highlighted shortcomings, and thus contribute to a more accurate interpretation of the results.

1. Introduction

> "*Regarde de tous tes yeux, regarde !*" ["Look with all your eyes, look!"]
> — Jules Verne, Michel Strogoff, 1876.

Many environmental processes are prone to significant variability, multiple couplings, feedbacks and stochastic forcings, resulting in high complexity that often jeopardizes the predictive capacity of deterministic, physically-based models (Wainwright and Mulligan, 2004; Young et al., 1996). Additionally, environmental measurement, in the broadest sense – monitoring, sampling, analysis –, often entails important challenges: significant costs, time constraints and inherent non-controlled conditions of *in situ* measurements may result in scarce and/or low-quality data with few replicates, despite recent developments in high-frequency monitoring strategies (Razguliaev et al., 2024; Smits et al., 2025). This may help understand the widespread use of empirical approaches in environmental studies, in order to relate two (or more) quantities that happen to co-fluctuate. In general, correlation- and regression-based techniques aim to go beyond, and possibly consolidate, the (typically graphical) evidence of a certain association between variables, with different objectives: from the quantification of the degree of association (*e.g.* through a correlation coefficient) to the development of a predictive model. A series of methods – which, like any statistical model, are underpinned by their own fundamental assumptions – have been developed for the foregoing purposes, and are now widespread and implemented in most existing tools for data analysis.

The genesis of this paper stems from a concerning observation exposed in the initial section: in peer-reviewed articles published over the last two decades in environmental sciences, there appear to be increasing cases of misuse of such techniques, which disregard their mathematical properties and underlying assumptions. In so doing, the main risk is to draw potentially erroneous conclusions – *e.g.*, to seemingly identify a "significant" association between two independent variables – from the application of an inadequate approach as a methodological backbone. The problem is not new, nor is its recognition, and pleas for an enlightened use of regression can be found in works such as Deming (1943), Tomassone et al. (1983), Helsel and Hirsch (2002) and Berthouex and Brown (2002). What is novel, however, is the unprecedented scale of the issue – exacerbated by the recent massification of data combined with the development of easy-to-run software for data analysis – ironically associated to the growing idea that correlation- and regression-based techniques are "basic statistics". As summarized by Mikkonen et al. (2019), "this simplicity is deceptive as the line-fitting

* Corresponding author.
E-mail address: damien.tedoldi@insa-lyon.fr (D. Tedoldi).

https://doi.org/10.1016/j.envsoft.2025.106526
Received 15 April 2025; Received in revised form 11 May 2025; Accepted 16 May 2025
Available online 19 May 2025

procedure is actually quite a complex problem".

Obviously, the intention of this paper is not to stigmatize specific studies by pointing out their shortcomings. Our main objective is rather to raise awareness among environmental scientists – whether they are contributing as article authors or acting as reviewers – about the inherent limitations of commonly applied correlation- and regression-based techniques, to outline key caveats, and to discuss possible ways to overcome the encountered limitations. Following an initial part that sets out the scope and scale of the problem, the article is structured around a series of fundamental assertions, illustrated by several examples and detailed by a few theoretical considerations. The purpose of these theoretical elements is not to compile yet another volume on statistics, but to offer the reader the analytical means to precisely grasp the problematic nature of the practices examined herein as well as the relevance of the proposed improvements.

2. Problem statement

To outline the addressed issue, we have first compiled eight graphical examples of statistical shortfalls published between 2017 and 2023 in well-established journals in environmental sciences (Fig. 1): these graphs were redrawn to ensure anonymity, but we keep their references available for skeptical readers.

The problems embodied by these graphs may be categorized as follows.

1. Application of a linear regression and/or calculation of a Pearson correlation coefficient for data points that visibly show a nonlinear pattern (Fig. 1A);
2. Presence of an influential point which alone is responsible for the modeled "trend" as well as its "statistical significance" (Fig. 1B and C) – even if the rest of the data points clearly indicate an absence of trend (Fig. 1B);
3. Violation of the normality assumption for correlation testing (Fig. 1B, C and G);
4. Violation of fundamental assumptions on regression residuals: homoscedasticity (Fig. 1D) and normality (Fig. 1E);
5. Pooling of data points that belong to different populations (Fig. 1F, and presumably Fig. 1G);
6. Extrapolation of a linear model outside of the range of available data (Fig. 1F and G) – even reaching non-physical values (Fig. 1F);
7. Interpolation of isolated clusters of data points, without any evidence of the intermediate trend (Fig. 1C and H).

Most importantly, they were treated as "positive" results (as reflected by the presence of a p-value on many of the graphs or in their caption), *i.e.* the authors concluded that there is a relevant association between the two variables under study, and/or that a linear trend is appropriate to describe their relationship. The fact that these glaring shortcomings did not elicit sufficient objection on the part of co-authors, editors, and reviewers, to prevent their publication, is already a relatively disturbing observation.

To demonstrate that the problem extends far beyond this set of eight studies, while avoiding confirmation bias, a systematic literature search was carried out. The topic of microplastics was selected as a currently very active scientific field, and this keyword was searched on Web of Science in association with the keywords "correlation" or "influencing factor" (the search was done between September and October 2024). Each article, in order of appearance according to the "relevance" sorting criterion, from a journal with an impact factor greater than 2.5, was scanned to check that the content or supplementary material did present a correlation or regression analysis. The search was stopped at 100 articles meeting these criteria: the earliest and median publication dates were 2017 and 2023, respectively; the median impact factor (in 2024) of the journals from which they were extracted was 7.6. These 100 articles were then meticulously examined to check whether the statistical methods are appropriate, or whether they exhibit one or several of the above-mentioned flaws.

The number of publications without any identified problem relating to correlations and regressions amounts to 25 out of 100. Conversely, 55 articles present *at least* one patent shortcoming, the most frequent being: linear regression/calculation of Pearson correlation coefficient on data points with no linear pattern (23), "statistical significance" due to the presence of an influential point (20), inference in linear regression with non-Gaussian or heteroscedastic residuals (13), and extrapolation or interpolation issues (13). In the remaining 20 articles, the details provided do not allow a definite assessment of the appropriateness of the methodology. Typical examples of this situation include the following: the nature of the coefficients presented in correlation matrices – Pearson, Spearman, or any other – is not specified; no normality test is mentioned prior to testing the significance of Pearson correlation coefficient, and the data are not available; or the sample size on which a correlation test was performed is not specified, while the data include repeated measurements on the same sites. A more subtle shortcoming, not quantified here, pertains to publications that merely comment on whether or not a correlation coefficient is "significant", without any consideration for the value of the coefficient itself.

We should certainly be alerted by the fact that more than half, and potentially up to 3/4, of the "most relevant articles" on microplastics and their links to other environmental factors, have based their conclusions on flawed statistical methods. In particular, this shows that the existence and availability of statistics manuals is not a sufficient bulwark against these kinds of misleading practices. Hence, with this position paper, we hope to contribute to a more informed use of empirical modelling among environmental scientists and, in turn, to help reduce the prevalence of statistical errors in published research.

In the subsequent developments, explanations will be illustrated using examples generated with random sequence simulators, designed to reflect actual configurations encountered in published literature. The reader will note that these examples are representative of the patterns shown in Fig. 1 and have not been artificially exaggerated. Synthetic data have the pedagogical advantage of showing that recreating datasets such as those depicted in Fig. 1 involves the violation of at least one of the statistical assumptions underpinning usual correlation analysis and/or linear regression. Another advantage is reproducibility: for readers willing to replicate the figures and calculations presented below, R scripts (R Core Team, 2021) are provided in the supplementary material accompanying this article.

3. Influential points

3.1. Quadratic and variance-based criteria are sensitive to influential points

Consider the following two examples, illustrated on Fig. 2.

Example A

X and Y are independent random variables. Let us assume that under usual environmental conditions, each of them follows a normal distribution, with a mean value of $\mu_X = 5$ and $\mu_Y = 10$, and a standard deviation of $\sigma_X = 2$ and $\sigma_Y = 3$, respectively. Evidently, the nature and characteristics of these distributions are unknown to the experimenter.

Joint measurements of X and Y under standard conditions lead to samples x_0 and y_0 with a sample size $n = 15$. Then an additional observation shows a large deviation from the common response of X and Y, both of which are significantly increased: $x_1 = 20$ and $y_1 = 30$. A discerning observer would readily recognize that this new point does not belong to the same underlying population. Nevertheless, as in many examples mentioned in Section 2, we consider a situation where the data are treated as such in order to explore potential correlation/trend, using the Pearson correlation coefficient and linear regression.

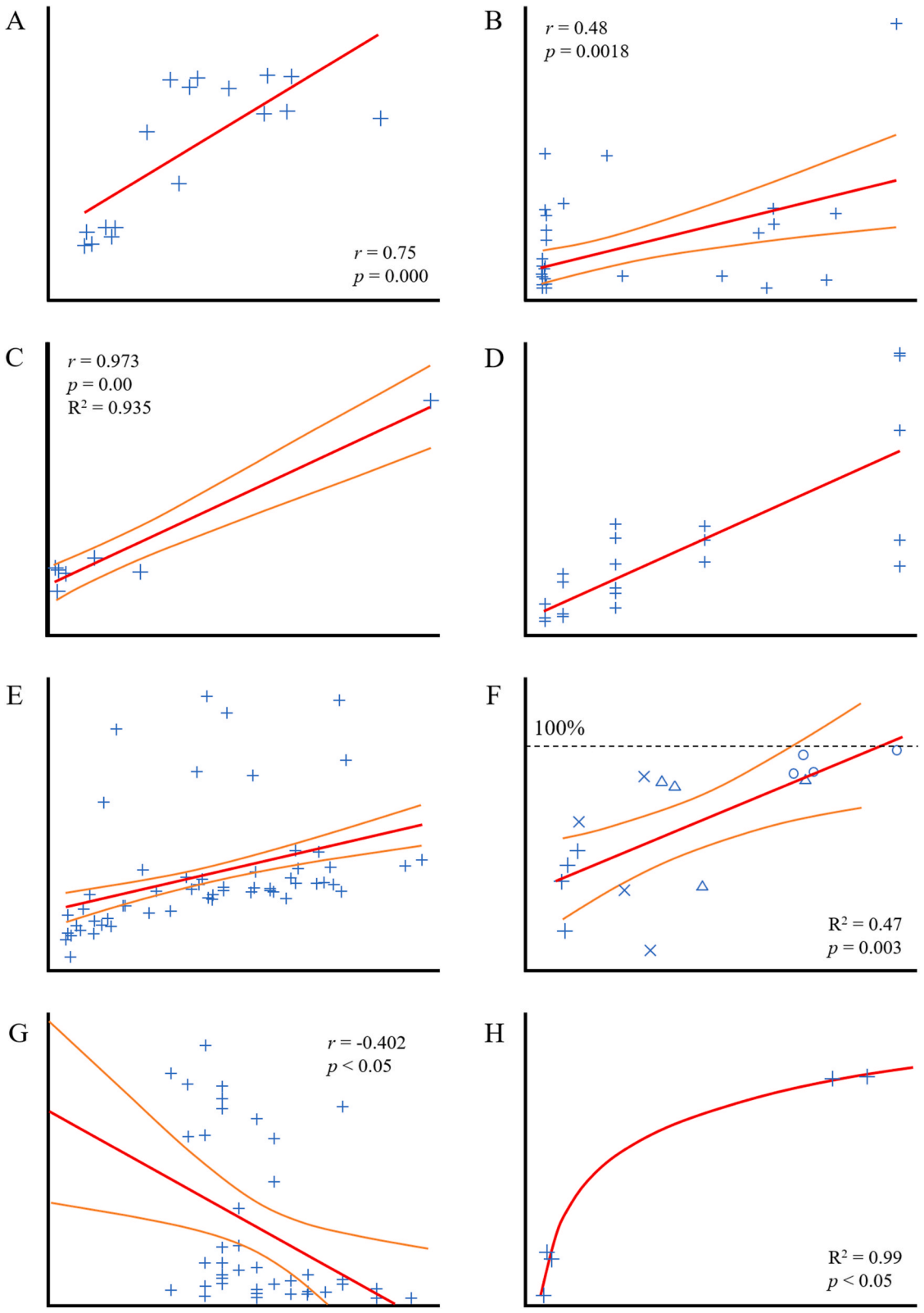


Fig. 1. Examples of misleading use of correlation- or regression-based techniques taken from scientific literature (A: Journal of Hazardous Materials; B: Journal of Exposure Science & Environmental Epidemiology; C: Environmental Pollution; D: Nature Communications Earth & Environment; E: Science of the Total Environment; F: Frontiers in Environmental Science; G: Environmental Science and Pollution Research; H: Environmental Science & Technology). To ensure anonymity, these figures have been redrawn and the legends and labels of the axes have not been included. The symbols correspond to the data points, the red line/curve to the least-squares regression model, and the orange curves, where present, to the 95 % confidence intervals on the regression line. On figure F, the different symbols correspond to different populations. When included on the original figures or in their captions, the notations r, p, and R^2 refer respectively to the Pearson correlation coefficient, the p-value associated with the test of significance of this correlation, and the coefficient of determination of the linear regression model.

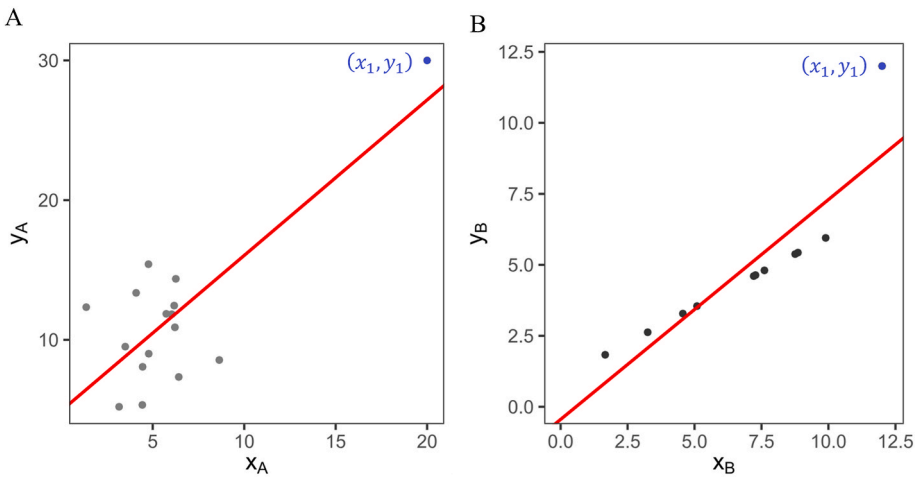


Fig. 2. Graphical illustration of Example A (left) and Example B (right). In Example A, while by construction there is no association between x_0 and y_0 (Pearson's $r = 0.05$ for the data points above), the addition of one point to the data series results in a much higher, and seemingly "significant", Pearson correlation coefficient ($r = 0.79$), as well as a positive slope of the regression line. The problem is different in Example B, where the two variables do have a linear association within a certain range, but the additional measurement results in a regression line that matches neither the linear portion of the data points nor the entire data series.

Example B

Consider two variables X and Y. In the usual range of variation of X, the two variables are linked by a deterministic linear relationship, whose coefficients are unknown to the experimenter. In this example, X is usually in $[0, 10]$, in which it follows a uniform distribution and $Y = 0.5X + 1$.

In order to characterize the relationship between X and Y, 10 joint measurements (x_0 and y_0) have been carried out in standard conditions, and an additional measurement has been taken outside this range, leading to $x_1 = 12$ and $y_1 = 12$. Linear regression is then applied to the whole dataset.

In both situations, the point (x_1, y_1) is called an *influential point*, since it has disproportionate effects on the correlation/regression results. This influence directly derives from the mathematical properties of usual correlation- and regression-based techniques. Generally, the indicator that is estimated by default when calling a "correlation" function on a statistical software is the Pearson correlation coefficient, r_{XY}. Readers will likely be familiar with the formula used to estimate this coefficient from n joint observations of X and Y:

$$r = \frac{\sum_{i=1}^{n} \left(x_{obs,i} - \bar{x}\right) \cdot \left(y_{obs,i} - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n} \left(x_{obs,i} - \bar{x}\right)^2 \cdot \sum_{i=1}^{n} \left(y_{obs,i} - \bar{y}\right)^2}} \tag{1}$$

where $x_{obs,i}$ (resp. $y_{obs,i}$) is the i^{th} observed value of the variable X (resp. Y), and $\bar{x}$ and $\bar{y}$ are their respective mean values. Less commonly acknowledged is apparently the sensitivity of this measure to values that deviate significantly from the center of the distribution. By construction, r tends towards $\frac{(x_1 - \bar{x}) \cdot (y_1 - \bar{y})}{|x_1 - \bar{x}| \cdot |y_1 - \bar{y}|} = \pm 1$ as the influential point (x_1, y_1) moves further away from the rest of the data points in both directions (like in Example A).

A similar effect occurs when computing the a and b coefficients of a regression line as *least-squares* estimates, which minimize the following criterion:

$$\sum_{i=1}^{n} \left(y_{obs,i} - (a x_{obs,i} + b)\right)^2 \tag{2}$$

Squaring the residuals inherently magnifies large differences, and therefore tends to discard models that deviate substantially from one or a few influential point(s). Hence, in Example B, the regression line displayed on Fig. 2B is considered a "better fit" than the line $Y = 0.5X + 1$ (*i.e.* the truly linear trend observed in the interval $[0, 10]$), which is penalized by the squared difference $(y_1 - (0.5x_1 + 1))^2$.

In many cases, such as those shown on Fig. 1B-C and Fig. 2A-B, preliminary visual assessment should be quite sufficient to identify influential points. This may be complemented by usual outlier detection techniques to reinforce the conviction that these points have likely been generated by different mechanisms, and therefore shall not be modeled with the same equation as the rest of the population. For more ambiguous settings, a helpful and probably underutilized indicator to pinpoint (potentially) influential points is the "leverage" that derives from the projection matrix. The latter quantifies the influence of each observation $y_{obs,j}$ on the predicted value $y_{sim,i}$, and leverage scores correspond to the diagonal elements:

$$\text{Leverage}_i = \frac{\partial y_{sim,i}}{\partial y_{obs,i}} \in [0, 1] \tag{3}$$

A value close to 1 for the i^{th} leverage score means that a variation in $y_{obs,i}$ will result in an identically large variation in $y_{sim,i}$, or graphically, that the regression line "sticks" to the i^{th} data point. A point with a high leverage score is not necessarily an influential point, as it may just be distant from the rest of the data points but aligned with them. However, it should raise the modeler's vigilance regarding its potential effect on the regression model, particularly if it is subject to some uncertainty. On R software, leverage scores are calculated using the *hatvalues* function.

An even more direct way of identifying influential points is provided by "leave-one-out" procedures, which consist of (i) omitting one observation among the n data points, (ii) applying a linear regression to the subset of remaining $n - 1$ points, and (iii) examining the change in the fitted parameters and/or in the predicted value for the omitted point. These steps are repeated n times (excluding all data points one at a time), thus highlighting – if any – the points whose absence would result in a significantly different regression model.

By way of illustration, these two methods have been applied to Example A (see supplementary material). The leverage score for all points drawn from the two independent Gaussian distributions, *i.e.*, x_0 and y_0, remains below 0.15 (median: 0.07), while it is as high as 0.85 for the point (x_1, y_1). Similarly, the absence of one point from the first series of data would only marginally alter the slope of the regression model (from 1.11 to a minimum of 1.07 and a maximum of 1.24), which would however drop to <0.1 if the influential point were left out.

3.2. The criteria used for evaluating the distance between observed and simulated values may have a major influence on the fitted relationship

Prior to any statistical consideration, regression can first be regarded as a calibration procedure, in which the model parameters are fitted to reach a minimum distance between observed and simulated values. That being said, it should be kept in mind that there are different ways to define a "distance" between two sets of values. As presented in the foregoing paragraph, usual applications of linear and nonlinear regressions are based on the least-squares method (Eq. (2)), in which the distance criterion is defined as the sum of squares of the residuals, *i.e.* the usual Euclidean or ℓ^2 norm:

$$Dist_{\ell^2}(y_{obs}, y_{sim}) = \sum_{i=1}^{n} \left(y_{obs,i} - y_{sim,i}\right)^2 \quad (4)$$

Conversely, a criterion based on the ℓ^1 norm considers absolute values of the residuals and thus mitigates the contribution of large differences:

$$Dist_{\ell^1}(y_{obs}, y_{sim}) = \sum_{i=1}^{n} \left|y_{obs,i} - y_{sim,i}\right| \quad (5)$$

Whatever distance criterion is chosen, it is essential to recognize that, when considering the *vertical* distances to the regression line/curve ($y_{obs,i} - y_{sim,i}$), all error is assumed to reside in Y, while X is assumed to be known without (or with negligible) uncertainty (Cantrell, 2008). This is, for example, appropriate when considering the calibration curve of any measurement device, *i.e.*, the output delivered by the device when it is submitted to known x values – ideally certified standards (Benisch et al., 2021). However, in the presence of uncertainties affecting both the X and Y variables, it has been shown that the slope of the regression line is underestimated (Francq and Govaerts, 2014; Mikkonen et al., 2019).

It can be easily conceived that the minimum of these two distance functions is not necessarily reached for the same set of model parameters. To demonstrate how different the results may be, let us go back to Example B. Fig. 3 compares the "optimal" linear and quadratic models derived from the minimization of the $Dist_{\ell^1}$ and $Dist_{\ell^2}$ criteria. The influential point for the least-squares estimates of the regression coefficients has, conversely, no effects on the coefficients fitted with the ℓ^1 norm: the latter do coincide with the "true" coefficients of the relationship between X and Y in the interval $[0, 10]$, even for a quadratic regression model, for which the best-fit coefficient of x^2 is almost zero.

This brings us to a fundamental statement: the "optimal" parameters of a model fitted to a data series are not intrinsic, they are a direct consequence of the distance criterion used, which should not be regarded as a "default" or "universal" option. The usual choice of the ℓ^2 norm has several theoretical advantages, provided that statistical assumptions on the regression residuals can be guaranteed (Section 4.2). For instance, if X and Y are indeed linearly related in the whole population, the least-squares estimates of the regression coefficients are unbiased. More generally, the calculations leading to, *e.g.*, confidence and prediction intervals of the regression, are not compatible with estimates derived from the ℓ^1 norm. However, it is certainly preferable to limit the ambitions of the regression approach while still achieving a reasonably accurate adjustment, rather than applying a flawed method, notably due to influential points – but more widely to a whole set of assumptions that will be shown to be actually quite restrictive, and often violated.

4. Regression as a modelling approach

4.1. Regression models aim at making predictions, not at drawing a line on a graph

Thus far, the described objective has been to achieve the closest fit to the observed data points by adjusting the parameters of a given analytical regression function. Things fundamentally change when regressions are regarded as a modelling approach. Then the aim is not just to *approximate* the available data, but to claim that the fitted relationship would still be valid for a new measurement – in other words, to *predict* the value of Y corresponding to any x value. As should be the case for all models, this objective goes hand in hand with an assessment of the magnitude of uncertainty. This broadened purpose leads to the introduction of a probabilistic framework, in which (i) Y is considered as a random variable, (ii) the whole regression model characterizes the conditional probability distribution $Y|\{X = x\}$, (iii) the regression function corresponds to the deterministic component of the model, and (iv) the residuals are treated as realizations of a random variable ϵ, which encompasses all sources of variability in Y that are not captured by the latter. Under the common approach of least-squares minimization, the regression function is an estimate of the conditional expectation $\mathbb{E}[Y|X = x]$, and the residuals are assumed to be Gaussian with specific properties (see Section 4.2).

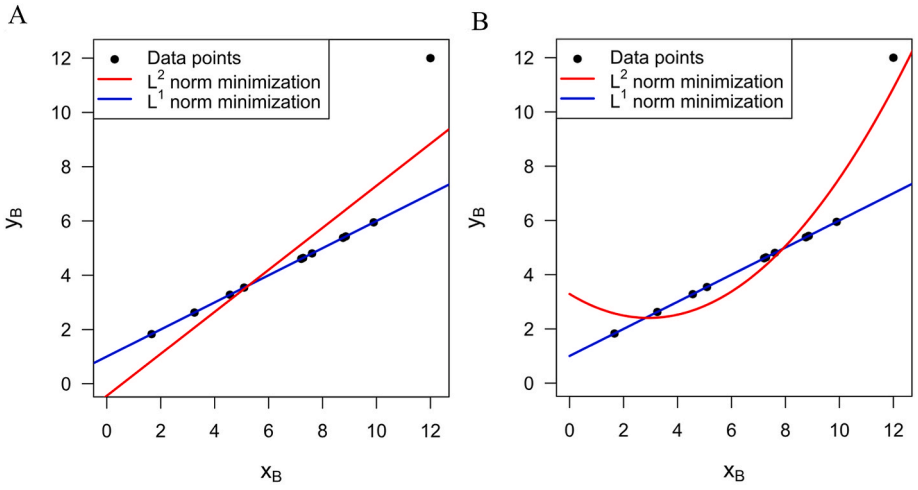


Fig. 3. Data points from Example B (10 measurements in a range where X and Y are linked by a linear relationship, and an additional point outside), along with the "optimal" linear (left) and quadratic (right) models fitted to the data by minimizing the $Dist_{\ell^1}$ (blue) and $Dist_{\ell^2}$ (red) criteria. The models adjusted with the $Dist_{\ell^2}$ criterion are shifted towards the influential point (the parabola naturally passes closer to it than the regression line due to an additional degree of freedom), but do not reproduce the linear trend between X and Y in the interval $[0, 10]$. Conversely, the "best fit" models obtained from the minimization of the $Dist_{\ell^1}$ criterion do match this linear trend and are not affected by the additional point.