Received: 2 November 2012, Revised: 10 April 2013, Accepted: 10 April 2013, Published online in Wiley Online Library: 12 July 2013

(wileyonlinelibrary.com) DOI: 10.1002/env.2213

When are two pieces better than one: fitting and testing OLS and RMA regressions[†]

Jane Friedman[a*], Andrew J. Bohonak[b] and Richard A. Levine[c]

This paper describes a new method of choosing between a simple linear and a two-phase linear reduced major axis regression model. For continuous two-phase linear models, differential evolution is used to estimate the parameters. The method can be used with either ordinary least squares or reduced major axis regression. There are no other methods in the literature for fitting two-phase reduced major axis regressions. Our new method is applied to problems in zebrafish kinetics bedload transport and a classic linear model of cricket chirping. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: differential evolution; model selection; reduced major axis regression; two-phase linear models; two-piece regression

1. INTRODUCTION

There are many biological and physical phenomena for which a two-piece linear function provides a good functional and statistical model. In these cases the location of the transition point is often of particular interest, because it reflects a qualitative change in the underlying phenomenon. For example, Danos and Lauder (2007) present a study of the spontaneous routine turns performed by zebrafish *Danio rerio* as part of their normal behavior. Danos and Lauder (2007) are interested in the relationship between kinematic properties of these turns (duration, curvature, turn angle, and angular momentum) and the fork length of the fish (length of the fish from the snout to the end of the middle caudal ray). As zebrafish grow, they pass through different ontogenetic stages (larval, juvenile, and adult) and undergo morphological changes. By looking at which kinematic variables have a simple linear relationship with fork length and which are better represented by a continuous piecewise linear model, the relationship between morphological development and kinematic behavior can be better understood.

Although methods have been developed to fit piecewise linear ordinary least squares (OLS) models to data, we are unaware of any work in the literature or software that addresses the fit of piecewise linear models in reduced major axis regression (RMA). (RMA is also known as standardized major axis regression, geometric mean regression or least product regression.) In their influential text, Sokal and Rohlf (2001) suggest that RMA may be more appropriate in situations where both variables are subject to error and the two variables have different units of measurement. RMA, in contrast to OLS, treats both variables symmetrically, making no assumptions about dependence. This paper fills a gap in the literature by presenting a method of fitting piecewise linear RMA models to data.

The question of when, if ever, RMA is the appropriate method to use is controversial. The large number of papers addressing this issue have reached many different conclusions (e.g. Laws and Archie, 1981; Seim and Sæther, 1983; McArdle, 1988; Ludbrook, 2010, and references therein). The details of this dispute are outside the scope of this paper. However, if RMA is the appropriate method to fit a simple linear model, then under the same conditions RMA should be used when fitting a two-phase linear model. The purpose in the present work is not to contribute to the discussion of which type of regression to use, but to contribute a new method that can be used to fit piecewise linear two-phase models that works for both OLS and RMA.

Smith (2009, pg. 482) presents a particularly lucid discussion of this issue and recommends the use of RMA in the following situations:

1. It seems arbitrary which variable (X or Y) is the independent and which the dependent variable.
2. The objective is to determine some mutual codependent "law" underlying the relationship between X and Y.
3. The slope of the line will be used to determine whether X and Y are isometric or whether Y shows positive or negative allometry.

The zebrafish study of Danos and Lauder (2007) used two-piece OLS regression. Following Sokal and Rohlf (2001), these authors justify their use of OLS in regressing kinematic variables against fork length on the grounds that fork length measurement error is small compared with the measurement error in the kinematic variables. However, fork length and the kinematic variables may both depend on the ontogenic

* *Correspondence to: Jane Friedman, Department of Mathematics, University of San Diego, San Diego, CA, U.S.A. E-mail: janef@sandiego.edu*

a *Department of Mathematics, University of San Diego, San Diego, CA, U.S.A.*

b *Department of Biology, San Diego State University, San Diego, CA, U.S.A.*

c *Department of Mathematics and Statistics, San Diego State University, San Diego, CA, U.S.A.*

[†] This article is published in Environmetrics as part of a Special Issue in memory of George Casella and his contributions to environmental statistics; guest-edited by Linda J. Young, Department of Statistics and Institute of Food and Agricultural Sciences, University of Florida, Gainesville, Florida, USA.

stage of the fish, fitting Smith's case 2. This observation is noteworthy because, as shown later in this paper, piecewise RMA regression produces qualitatively different results for some of Danos and Lauder's variables than piecewise OLS regression.

We first describe in Section 2 the method we will use for fitting piecewise linear models. In Section 3, we evaluate the algorithms through a series of simulation studies. In Section 4, we apply the proposed methods to real data. Section 5 concludes with a discussion of future work.

2. PIECEWISE LINEAR REDUCED MAJOR AXIS REGRESSION

2.1. Reduced major axis regression, ordinary least squares, and two-phase linear models

In the RMA set-up, we consider two variables X and Y each measured with error, with standard deviations σ_X and σ_Y, respectively. The variables are related, with correlation coefficient ρ. Despite the traditional notation, no dependent–independent variable relationship is defined. In the early 20th century RMA first appears in the literature as a way of approaching regression in this situation. The first modern exposition appears in Clarke (1980). The reduced major axis line standardizes the variables and fits a line of slope $\beta = \pm\sigma_Y/\sigma_X$, (Sokal and Rohlf, 2001; pg. 544) which is the geometric mean of the lines regressing Y on X and X on Y. Barker *et al.* (1988) shows that this approach identifies the same line whether Y is regressed on X or X regressed on Y. Nonetheless, in the remainder of this section, we will take a least squares approach to the problem for build-up from OLS regression.

The primary goal of this paper is to provide tools for fitting two-phase linear models, whether the lines are fit by using RMA or OLS. By two-phase, we are presuming a breakpoint x_c. Separate regression lines are fit on the data $\{(x_i, y_i) : x_i \leqslant x_c, i = 1, \ldots, n\}$ and the points $\{(x_i, y_i) : x_i > x_c, i = 1, \ldots, n\}$.

Fitting the best continuous two-phase linear model, with constraints imposed on the location of the change point, is a nonlinear constrained optimization problem. In general, such problems lack closed form solutions. We present a method for approximating the best two-phase linear model and testing the hypothesis that the simple linear model is a better fit. To perform the approximation we use differential evolution (DE), described in Section 2.3. DE has been used for a wide range of constrained nonlinear optimization problems, with performance as good or better than alternative approaches (Vesterstrøm and Thomsen, 2004; Price, 1996). Our particular use of the method is not found in the literature. We use simulations to test the null hypothesis that the simple linear model is a better fit. This approach to model selection is related to the methods described in Gijbels and Goderniaux (2004) and Li *et al.* (2011). In this paper, the focus is on RMA, with minor adjustments for OLS. Both RMA and OLS fit a line to data that minimizes the deviations of the data values from the fitted line. Given n data points (x_i, y_i) and the model $Y_i = a + bX_i + \epsilon_i$, OLS minimizes

$$\sum_{i=1}^{n}\{y_i - (bx_i + a)\}^2 \tag{1}$$

with respect to a and b (Figure 1).

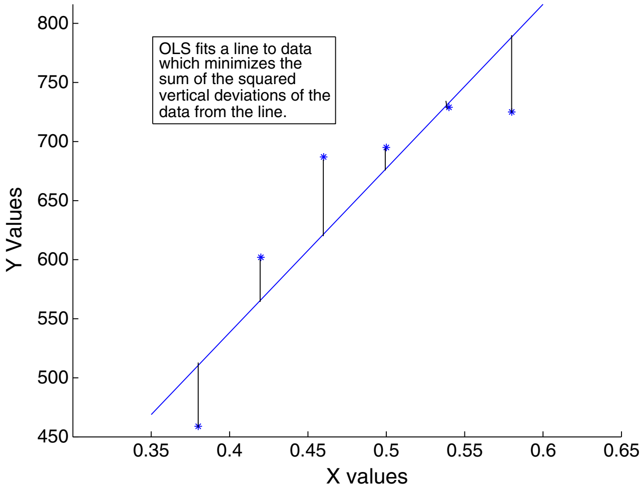

Figure 1. Illustration of ordinary least squares (OLS) regression line fitted to data for comparison with RMA regression in Figure 2

When there is no scientific reason to define one variable as dependent on the other in the usual regression sense, a method such as RMA may be more appropriate. RMA fits the line to data which minimizes the sum of the areas of the right triangles which have legs parallel to the x-axis, y-axis, and hypotenuse on the fitted line. Thus RMA minimizes

$$\sum_{i=1}^{n}\left|\{y_i - (bx_i + a)\}\left\{x_i - \left(\frac{y_i - a}{b}\right)\right\}\right| \tag{2}$$

with respect to a and b (Figure 2).

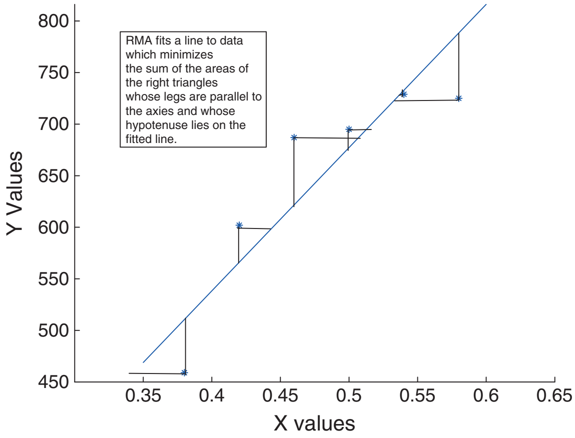

Figure 2. Reduced major axis regression regression line fitted to data

For both of these methods, there are well-known formulae for the slope and intercept of the fitted line. Suppose we have data $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$, and that the line of best fit is $\hat{Y} = \hat{a} + \hat{b}X$. Then for OLS

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

and for RMA

$$\hat{b} = \text{sign}\left(\sum_{i=1}^{n} x_i y_i\right) * \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} x_i^2}}. \tag{3}$$

In both cases the equation of the fitted line line is given by

$$Y = (\bar{Y} - \hat{b}\bar{X}) + \hat{b}X \tag{4}$$

so that the fitted line passes through $(\bar{X}, \bar{Y})$, where $\bar{X}$ is the mean of the x-values and $\bar{Y}$ is the mean of the y-values.

Figure 2 may remind readers of the discussion of orthogonal regression in Casella and Berger (2002, Section 12.2). The two methods are similar in that both entail a "least-triangles" approach to estimation. Orthogonal regression, as the name implies, takes an OLS approach minimizing orthogonal projections onto the best fit line. The right triangles in orthogonal regression thus have the 90 degree angle at the best fit line, as opposed to reduced major axis regression where the best fit line is the hypotenuse of the minimized right triangle. Nonetheless, both methods present as error in variables models. However, unlike more traditional applications of measurement error models, neither method requires a strictly defined relationship of dependent variable regressed on independent variable. Carr (2012) presents an accessible exposition of the two methods with an application to geology. In this article we focus on RMA.

2.2. Fitting two-phase piecewise linear models

We have developed a numerical method to test the null hypothesis that a simple linear model (either OLS or RMA regression) fits the data against the alternative hypothesis of a two-phase linear model (which may be required to be continuous) and to estimate the best two-phase linear model parameters. The algorithm may be found in the Appendix.

First, the best simple linear model and the best two-phase linear model are found or, in the case of the continuous model, estimated. For each of these, the sum of squared errors is computed. Let $\Delta\epsilon$ be the difference in these errors. $\Delta\epsilon$ measures the improvement in fit of the two-phase model over the linear model. We wish to test the null hypothesis that the linear model is a better fit, taking into account the increase in parameters. Data are simulated N times according to the simple linear model and the process is repeated, fitting or approximating the best two-phase linear model with $\Delta\epsilon_i$ the improvement in fit for the i^{th} replicant data set, $i = 1, \ldots, N$. An empirical p-value can be calculated as the proportion of $(1+\rho)\Delta\epsilon_j > \Delta\epsilon,\ j = 1, \ldots, N$, where $\rho \geqslant 0$ is a tuning parameter. One would accept the null hypothesis if this p-value is larger than some significance level α.

To fit a two-phase linear model which is not constrained to be continuous is simpler than fitting a continuous two-phase linear model. In the discontinuous case, because all of the information is contained in the data, one may assume that one of the data points is the change point. Thus, it is only necessary to test each x-value in turn as the breakpoint and see which one produces a model with minimal error.

To test the null hypothesis, we wish to produce "replicant data sets", data simulated to fit the best simple linear model for the data. For OLS, this is straightforward. Each data point corresponds in an obvious way to the unique point on the fitted line with the same x-coordinate. Data are simulated by taking these points and bootstrapping the residuals (see the Appendix).

The question of how to produce replicant data sets for RMA is nontrivial. Although RMA minimizes the sum of areas of triangles, there is no unique triangle with a given area. There is also no obvious unique point on the fitted line corresponding to each data point. In our implementation, we consider the point where the altitude of the minimized triangle intersects the fitted line to correspond to the data point. If the fitted line has equation $y = a + bx$, then

$$(x_i^*, y_i^*) = \left(\frac{-ab + y_i b + x_i}{b^2 + 1}, \frac{y_i b^2 + x_i b + a}{b^2 + 1} \right)$$

is the point corresponding to the data point (x_i, y_i). To these points, we add the triangle altitudes as bootstrapped vectors.

If the alternative hypothesis of a two-phase model is accepted, then an empirical bootstrap confidence interval for the breakpoint can be constructed. Create M replicant data sets by bootstrapping the original data points (note this is different from the process described in the previous paragraph). For each of these replicant data sets, fit the best piecewise linear model. Thus, including the original value found, we have $M + 1$ values for the breakpoint. By sorting these and taking the $(\alpha/2)^{\text{th}}$ value through the $(1 - (\alpha/2))^{\text{th}}$ value, a $(1 - \alpha)\%$ confidence interval for the breakpoint is constructed.

In cases where the model is required to be continuous at the change point, it can no longer be assumed that the optimal solution has a change point which is equal to one of the data points. For example, the data $(1, 1), (2, 2), (3, 3), (6, 4), (7, 3), (8, 2)$, exactly fit on the lines $y = x$ and $y = -x + 10$ with a change point at (5,5). In the case of OLS, the literature does contain methods for approximating the solution, (e.g., Hudson, 1966; Lerman, 1980; Muggeo, 2003; Vieth, 1989). Because no such method has been developed for RMA, the method presented here is noteworthy.

Differential evolution, a method of searching a continuous parameter space for the solution to a constrained nonlinear optimization problem, can be used to approximate the parameters in the best continuous two-phase linear model for either OLS or RMA regression. Because fitting a continuous two-phase linear model is a nonlinear optimization problem, some numerical method must be used. We chose DE because it has shown good performance for a wide range of problems, and indeed it works well here. See Storn and Price (1997a), Storn and Price (1997b), Feoktistov (2006), Price *et al.* (2005) and Das and Suganthan (2011) for more information on DE.

2.3. Differential evolution

Differential evolution is a stochastic direct search method for global optimization over a continuous parameter space. DE searches a D-dimensional parameter space for a vector of parameters to minimize a cost function, which can be nonlinear and non-differentiable. DE was designed to have the following desirable properties (Storn and Price, 1997a; pg. 342):

(1) Ability to handle non-differentiable, nonlinear, and multimodal cost functions.
(2) Parallelizabilty to cope with computationally intensive cost functions.
(3) Ease of use, that is, few control variables to steer the minimization. These variables should be robust and easy to choose.
(4) Good convergence properties, that is, consistent convergence to the global minimum in consecutive independent trials.

For continuous two-phase regression, the parameter space has $D = 4$ dimensions: the two coordinates of the change point, and the two slopes of the lines. The intercepts of the lines are determined by the requirement that both lines pass through the change point. The cost function to be minimized is the sum in (2) (for OLS, minimize the sum in (1)). At each step or generation, G, of the algorithm there is a current population of N D-dimensional vectors:

$$x_{k,G}, \quad k = 1, 2, \ldots, N.$$

Following Storn and Price (1997b), we take N to be $10 * D = 40$. The first generation is chosen randomly. New trial vectors are produced as candidates for inclusion in the next generation by using two operations called mutation and crossover.

To add mutation to the algorithm, a weighted difference of two population vectors is added to an existing population vector. Each vector $x_{k,G}$ in the current generation is mutated producing the vector

$$v_{k,G+1} = x_{r_1,G} + F \cdot (x_{r_2,G} - x_{r_3,G}) \tag{5}$$

where r_1, r_2, r_3 are unequal random integers chosen from $[1, N]$, and F is chosen randomly from $(0.5, 1)$.

Crossover is applied to the mutated vectors to create more diverse trial vectors. Form the vector $u_{k,G+1} = (u_{1,k,G+1}, u_{2,k,G+1}, \ldots, u_{D,k,G+1})$, where

$$u_{l,k,G+1} = \begin{cases} v_{l,k,G+1} & \text{if } \texttt{rand}(\text{l}) \leqslant CR \text{ or } l = \texttt{randint}(k) \\ x_{l,k,G} & \text{otherwise} \end{cases} \tag{6}$$

where $\texttt{rand}(\text{l})$ is a uniform random number on $(0, 1)$, $\texttt{randint}(k)$ is randomly chosen from the set $\{1, 2, \ldots, D\}$ and $CR \in [0, 1]$ is a constant. A default value of $CR = 0.9$, works well, in the sense that an appropriate model will be selected in most cases. We constrain the coordinates of the change point. Although the default is to constrain the change point to lie within the range of the data points, it may be advisable to put tighter bounds on the location of the change point (for example, restricting it to the middle 90% of the data).

The costs of the trial vector $u_{k,G+1}$ and $x_{k,G}$ are compared. The vector with the smaller cost becomes $x_{k,G+1}$. We continue in this manner until a chosen stopping criterion is satisfied. We use the stopping criterion recommended in Zielinski and Laur (2008): Compute `diff`, the difference between the maximum cost and minimum cost for all parameter vectors in the current generation. If `diff` is less than some predetermined value, then stop.

3. SIMULATION EXPERIMENTS

In the simulations in the later text, `maxit`, the maximum number of iterations allowed in any one implementation of DE was set at 1000 and `numsims`, the number of replicant data sets used to compute an empirical p-value, was set at 100. In our implementations of DE, unless otherwise stated, we use $\rho = 0$, $N = 40$ and CR$= 0.9$ and the stopping criterion is `diff`$= 0.0001$. The location of the change point is restricted to the middle 90% of the data. The simulated data sets were analyzed using the algorithm described in Section 2.2, implemented in MATLAB on a MAC with a 2.5 GHz Intel Core i5 processor and 8GB of memory (see the supplementary website of this journal for the Matlab code).[‡]

3.1. Simulations of linear data

Fifty linear data sets were simulated for four values of s. Here s is the standard deviation of zero mean normal error added to both coordinates of each data point. Our new method was used to test the null hypothesis of no change point for all 200 simulations. Table 1 suggests that the method rejects the null hypothesis unacceptably often for the simulated data (at least 28% of the time for each value of s tested). However, in cases where the null was rejected, the linear model and the two-piece model were usually so similar that they predicted nearly equal values of Y throughout most of the range of X. This suggests that when the alternative hypothesis of two-phase regression is accepted, it should always be interpreted in the context of differences in slopes and intercepts between the two models. Future work on this problem may attempt to establish minimum thresholds for slope and intercept as a function of sample size.

Table 1. Results from simulation of linear data. Fifty data sets were simulated for each value of s. The columns include the standard deviation s considered, the maximum, minimum, average, and standard deviation of the p-values across the 50 simulated data sets, as well as the percentage of p-values above the 5% level

s	max p	min p	mean p	SD p	% $p \geqslant 0.05$
1	0.92	<0.02	0.28	0.28	78
5	0.97	<0.02	0.30	0.28	76
10	0.83	<0.02	0.32	0.26	82
12	0.91	<0.02	0.28	0.26	72

Figure 3 shows the two-phase linear model and linear model for one example of simulated linear data, with $s = 10$. The p-value found was zero. The slope and intercept for the simple linear model are 1.04 and -4.7, respectively. The two slopes for the two-phase linear model are -1.34 and 1.09, and the two intercepts are 17.58 and -8.14 with change-point at (10.52,3.35), which is close to the edge of the range of the data values.

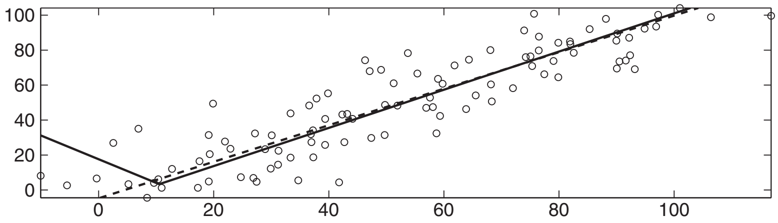

Figure 3. Illustrative example of the fit of simple linear and piecewise linear models for simulated linear data where the hypothesis of no change point is rejected

To further study this matter, we consider the prediction error over the simulated data sets. Table 2 reports results of 50 sets of simulated linear data for several values of s. For each of these replicant data sets, the average mean squared error for the linear model was calculated as was the average mean squared error for the change point model. The difference in these averages was also calculated. The table reports the average of these differences over 1) all fifty replicant data sets and then separately, 2) those in which the simple linear model was not rejected, and 3) those in which the simple linear model was rejected. It is apparent that the method generally selects the piecewise model for those situations where it is truly a better fit.

[‡] Supporting information may be found in the online version of this article.

Table 2. Results from simulation of linear data. The columns include the standard deviation s considered. For each value of s there are also columns for the overall mean of the difference between mean squared errors for the two models and the mean separately for those data sets where $p < 0.05$ and the percentage of p-values above the 5% level. Also reported is percent of $p \geq 0.05$

s	overall mean diff mse	mean diff mse $p < 0.05$	mean diff mse $p \geqslant 0.05$	$\% p \geqslant 0.05$
1	0.04	0.09	0.03	76
5	1.90	3.31	0.96	60
10	7.62	16.74	4.74	76
12	12.40	26.38	7.99	76

Table 3. Results from simulation of continuous two-phase linear data. The columns present the standard deviation s considered and the average and standard deviation of the change point x_c and slopes for each linear piece, denoted $sl1$ and $sl2$

s	mean x_c	SD x_c	mean $sl1$	SD $sl1$	mean $sl2$	SD $sl2$
1	50.49	0.37	1.01	0.02	9.99	0.10
5	49.85	1.93	1.09	0.11	9.66	0.42
10	47.75	3.59	1.29	0.28	9.97	0.10
12	46.55	5.17	1.37	0.44	8.17	0.73

Table 4. Results from simulation of continuous piecewise linear data, with different values for N and CR, each row represents the values from 50 replicant data sets. In all cases $s = 5$, $\rho = 0$. The columns present the values of N and CR and the average and standard deviation of the change point x_c and slopes for each linear piece, denoted $sl1$ and $sl2$, the empirical p-value, and the time t in seconds

N	CR	mean x_c	SD x_c	mean $sl1$	SD $sl1$	mean $sl2$	SD $sl2$	% $p \geq .05$	mean t	SD t
40	0.9	49.13	1.97	1.07	0.14	9.67	0.54	0	192	7
40	0.95	48.93	1.94	1.08	0.13	9.56	0.47	0	185	6
50	0.5	48.77	2.19	1.08	0.16	9.50	0.49	0	309	9
50	0.7	48.84	2.35	1.06	0.10	9.56	0.51	0	280	25
50	0.9	49.43	2.03	1.12	0.17	9.67	0.49	0	276	7
50	0.95	48.76	1.77	1.07	0.12	9.49	0.43	0	302	18

3.2. Simulations of data with a change point

The method was also tested on simulated data with a change point and RMA regression. Points were chosen on two lines intersecting at the point (50,50): $y_i = x_i = i$, for $i = 1, \ldots, 50$, and then the points $x_i = i$, $y_i = 10x_i - 450$ for $i = 51, \ldots, 100$. Random normal variates with mean 0 and standard deviation s were added independently to the x-coordinate and y-coordinate values, as in the previous section. For each value of s the method was tested on 50 data sets. Again we test the null hypothesis of no change point. Here low p-values and values x_c close to 50 would indicate that the method was performing well. All of these simulations had a p-value of zero, except three for $s = 12$ (p =0.01, 0.01, 0.06). See Table 3.

Table 4 shows the results of simulations with different values of the DE parameters N and CR. Our results demonstrate that the default values used in these simulations ($N = 40$ and $CR = .9$) are good choices, both in terms of how well the parameters of the model were estimated and in terms of the time needed for analysis.

3.3. Results from simulations with a penalty tuning parameter

In the simulations described previously, the tuning parameter $\rho = 0$. The parameter ρ can be thought of as an additional penalty compensating for the increase in parameters in the piecewise model. A researcher can choose a value of ρ that takes into account the particulars of the application and the cost of incorrectly rejecting the null hypothesis of a simple linear model.

Table 5 shows the results for various values of ρ for both simulated linear and two-phase linear data. These data were generated in the same manner as described in Sections 3.1 and 3.2. In all cases $s = 10$, and the coordinates of the change point were restricted to the middle 90% of the data. For these simulations, taking $\rho = 0.5$ results in a decrease in the probability of a type I error with good power.

Table 5. Results from simulation with tuning parameter ρ. The columns present the ρ values considered and the maximum, minimum, average, and standard deviation of the p-values across the 50 simulated data sets, as well as the percentage of p-values above the 5% level

ρ	simple linear	max p	min p	mean p	SD p	% $p \geqslant 0.05$
0.3	yes	0.95	<0.02	0.36	0.30	88
0.5	yes	0.89	<0.02	0.46	0.28	96
0.7	yes	0.97	0.02	0.49	0.28	94
0.3	no	0.02	<0.02	0.001	0.004	0
0.5	no	0.03	<0.02	0.004	0.009	0
0.7	no	0.08	<0.02	0.01	0.02	8

4. EMPIRICAL TESTS

4.1. Parameters used

In all the examples in the later text, `maxit` is set to 1000, `numsims` is set to 100, $N = 40$, $CR = 0.9$, $\rho = 0$, and `diff` $= 0.0001$, unless otherwise stated. We define `bsims` as the number of simulated data sets used to construct a bootstrap confidence interval; unless otherwise stated `bsims` $= 1999$. Because the problem of fitting a discontinuous linear model is trivial in comparison to fitting a continuous model, the examples presented focus on the continuous case.

4.2. Zebrafish kinematics

Danos and Lauder (2007) describe the results of a study of the kinematics of routine turning in zebrafish *Danio rerio*. The fish in the study varied across the natural size and ontogenic range of the fish. Fish ranged in size (fork length) from 0.38 to 1.97 cm. In Danos and Lauder (2007), OLS regression was used and four different variables were regressed against fork length: turn angle, angular velocity, duration, and curvature. Their results are presented in Table 6.

Table 6. Four regression models fit in Danos and Lauder (2007), with change point and CI on that change point for the two-phase linear models

Response variable	Model type	Change point, x_c	CI on x_c
Turn angle	two-phase linear	1.18	(0.90,1.46)
Angular velocity	two-phase linear	1.16	(0.93,1.39)
Duration	linear	–	–
Curvature	linear	–	–

Table 7. Results from applying the methods in Section 2 with both RMA and OLS to the data analyzed in Danos and Lauder (2007). The columns include the p-value for testing the hypothesis of a linear model against the alternative of a two-phase model, the breakpoint and the confidence interval on the breakpoint, and the time in seconds to perform the analysis on a Mac with a 2.5 GHz Intel Core i5 processor and 8 GB of memory

Variable	Regression type	Model selected	p	x_c	95% CI for x_c	secs
Turn angle	RMA	two-phase linear	0.02	0.9	(0.46,1.24)	3267
Turn angle	OLS	linear	0.26	–	–	103
Angular velocity	RMA	two-phase linear	<0.01	1	(0.48,1.24)	3783
Angular velocity	OLS	two-phase linear	0.04	1.1	(0.80,1.37)	3818
Duration	RMA	two-phase linear	0.01	1.47	(0.48,1.94)	3198
Duration	OLS	linear	0.07	–	–	104
Curvature	RMA	linear	0.07	–	–	105
Curvature	OLS	linear	0.51	–	–	94

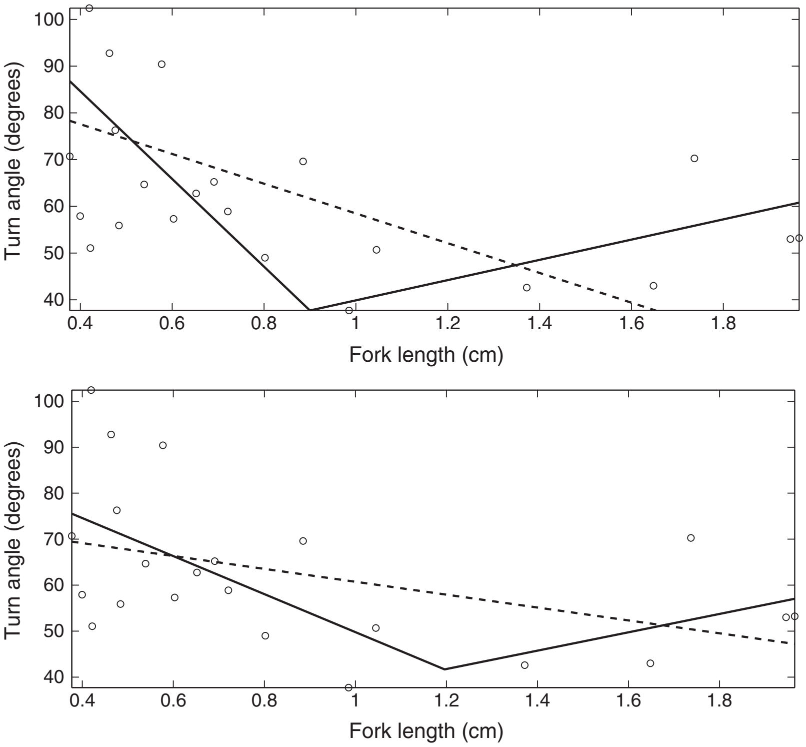

Figure 4. Simple linear (dashed) and continuous piecewise linear (solid line) models for the turn angle data. Reduced major axis models on top and ordinary least squares models below

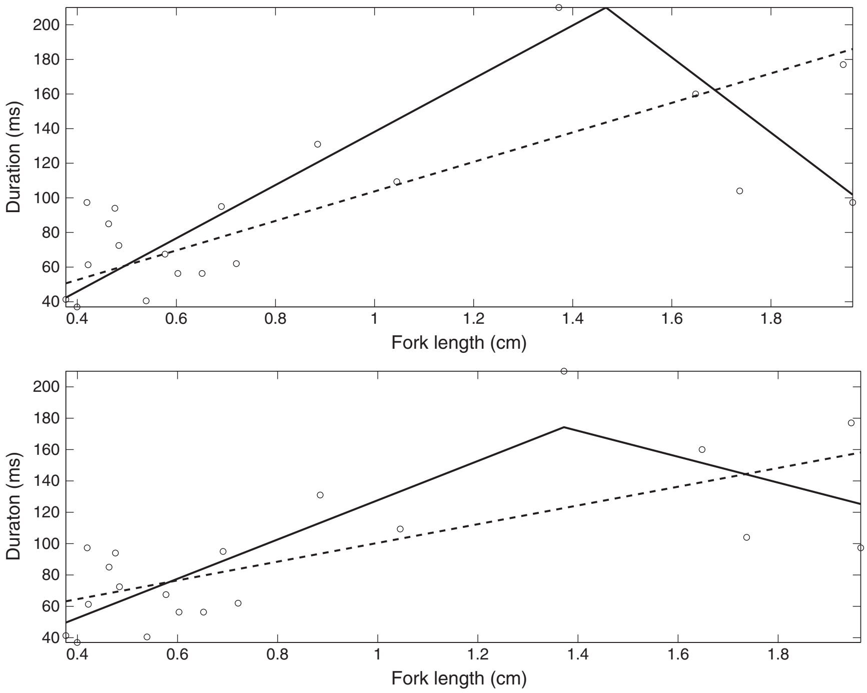

Figure 5. Simple linear (dashed) and continuous piecewise linear (solid line) models for the duration data. Reduced major axis models on top and ordinary least squares models below

The kinematic variables and fork length are both likely to depend on the ontogenic stage and age of the fish, which suggests that RMA may be more appropriate than OLS here. Our new method was was used to reanalyze the raw data of Danos and Lauder (2007), provided by

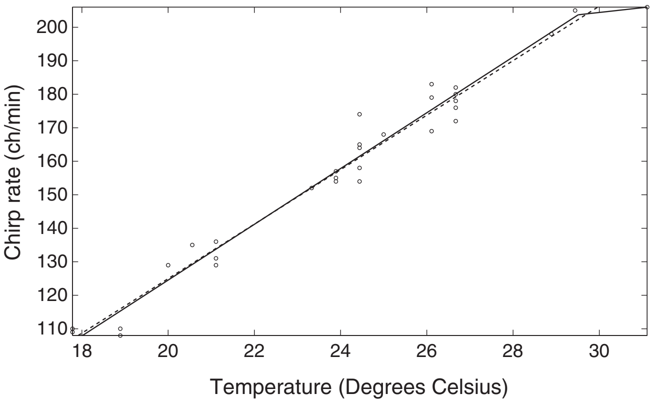

Figure 6. Best reduced major axis line and best continuous two-phase linear reduced major axis model for the cricket chirping data

Table 8. Reduced major axis (RMA) and ordinary least squares (OLS) results for Granite Creek bed load transport data

Type	Change point	95% CI for x_c	$\hat{b}_1$	$\hat{b}_2$	p-value	s
RMA	(6.04,0.05)	(5.50,11.16)	0.01	0.17	<0.01	3441
OLS	(4.83,0.01)	(4.05,11.10)	0.004	0.11	<0.01	1878

Nicole Danos (personal communication). Results for OLS and RMA are presented in Table 7. For two of the four variables (turn angle and duration), rejection of the linear model depends on which regression method is used (Figures 4 and 5). This result underscores the importance of providing a robust method that can be used with either RMA or OLS, leaving the choice of regression up to the practitioner.

4.3. Bedload transport

Bedload transport in coarse-grained channels is related to the flow or discharge level of the stream, but is also influenced by other factors including spatial and temporal variability in the sediment available for transport. Bedload transport is generally believed to occur in phases, and after a change point, higher flow rates are associated with substantially increased rates of bedload transport. Ryan and Porth (2007) and Ryan *et al.* (2002) suggest that piecewise linear regression can be used to determine the point at which this change in transport rate occurs. In Ryan and Porth (2007), a tutorial is presented for fitting a piecewise linear OLS model to bedload transport data, including raw data for Little Granite Creek (these data are analyzed below). Ryan and Porth first use a LOESS smoother to estimate the location of the change point, and then PROC NLIN in SAS to estimate the regression model parameters.

Our results for OLS are similar to those that Ryan and Porth present. They found the breakpoint to be 4.83 with a 95% confidence interval of (4.13,6.99), with slopes 0.0043 (first) and 0.1046 (second). Note that in Table 8 the 95% confidence interval for the breakpoint found using our methods with RMA does not include the breakpoint found using our methods with OLS.

4.4. Effect of temperature on cricket chirping

The method was also applied to a dataset whose biological basis suggests no change point should be present. The assumption that the rate of chirping in crickets is linearly dependent on temperature is so well-known that it is used as an example in introductory mathematics texts (see for example, Stewart, 2007; pg. 21). We applied our proposed piecewise RMA fitting routine to the data found in the online supplement to Walker and Collins (2010) on chirp rates for the cricket *Oecanthus fultoni*. We failed to reject the simple linear model ($p = 0.20$). Additional evidence in support of the simple linear model is provided by examining the best piecewise model found which is essentially the same as the simple linear model (Figure 6).

5. CONCLUSIONS AND FUTURE WORK

A number of different methods exist for fitting a line to data points. RMA is a commonly used method, recommended by some researchers for situations where OLS is not appropriate. This work presents a method for fitting two-phase continuous piecewise linear models to data by using differential evolution. The method can be applied using either OLS or RMA line fitting. It is particularly notable that this method can be used to fit continuous piecewise RMA models, because no such methods have previously appeared in the literature. The method uses differential evolution to approximate the best two-phase continuous piecewise linear model for the data, and calculates an empirical p-value for the null hypothesis that a simple linear model is a better fit. The method can also provide a bootstrap confidence interval for any one of the parameters of particular interest (often that will be the value of the x-coordinate variable at the change point).

The problem being solved here is a nonlinear constrained global optimization problem. The numerical approach in this paper is potentially costly to implement. In the future, we plan to explore parallel implementations of the DE routine, and its variants, as well as DE parameter choice toward improved computational efficiency. Additionally, as part of our future work, we will expand the method to fit models with multiple change points. This might be performed sequentially by first testing a one change point model against a no change point model, and if the null hypothesis of no change points is rejected testing the hypothesis of a two change point model against a one change point model.

Acknowledgements

We thank three anonymous reviewers for comments that lead to significant improvements in our paper. We also thank Special Issue Editor Professor Linda Young and Editor-in-Chief Professor Walt Piegorsch for their efforts in proposing and then expertly and efficiently putting this issue together in honor of George Casella. RL was in the first crop of environmental statistics training grant doctoral students under George and AB poured over Casella and Berger under his tutelage in STAT 409, both at Cornell in the early ’90s. Our paper, and we believe this issue, is the epitome of George’s approach of collaborative efforts to developing and applying novel statistical methods for scientific problem solving. A brilliant statistician, a great man, a devoted father, and a dear friend–you left us too soon.

REFERENCES

Barker F, Soh YC, Evans RJ. 1988. Properties of the geometric mean functional relationship. *Biometrics* **44**: 279–281.

Casella G, Berger R. 2002. *Statistical Inference*, (2nd edn). Duxbury: Pacific Grove, CA.

Carr J. 2012. Orthogonal regression: a teaching perspective. *International Journal of Mathematical Education in Science and Technology* **43**: 134–143.

Clarke MRB. 1980. The reduced major axis of a bivarite sample. *Biometrika* **67**: 441–446.

Danos N, Lauder G. 2007. The ontogeny of fin function during routine turns in zebrafish *Danio rerio*. *Journal of Experimental Biology* **210**: 3374–3386.

Das S, Suganthan PN. 2011. Differential Evolution: a survey of the state-of-the-art. *IEE Transactions on Evolutionary Computation* **15**: 4–31.

Feoktistov V. 2006. *Differential Evolution: In Search of Solutions*, Optimization and Its Applications no. 5. Springer Science and Business Media, LL: New York.

Gijbels I, Goderniaux AC. 2004. Bootstrap test for change-points in non-parametric regression. *Journal of Nonparametric Statistics* **16**: 591–611.

Hudson D. 1966. Fitting segmented curves whose join points have to be estimated. *Journal of the American Statistical Association* **61**: 1097–1129.

Laws EA, Archie JW. 1981. Appropriate use of regression analysis in marine biology. *Marine Biology* **65**: 13–16.

Lerman PM. 1980. Fitting segmented regression models by grid search. *Journal of the Royal Statistical Society Series C. Applied Statistics* **29**: 77–84.

Li C, Wei Y, Chappell R, He X. 2011. Bent line quantile regression with application to an allometric study of land mammals’ speed and mass. *Biometrics* **67**: 242–249.

Ludbrook J. 2010. Linear regression analysis for comparing two measurers or methods of measurement: but which regression? *Clinical and Experimental Pharmacology and Physiology* **37**: 692–699.

McArdle BH. 1988. The structural relationship in biology. *Canadian Journal of Zoology* **66**: 2329–2339.

Muggeo V. 2003. Estimating regression models with unknown break-points. *Statistics in Medicine* **22**: 3055–3071.

Price K, Storn R, Lampinen J. 2005. *Differential Evolution – A Practical Approach to Global Optimization*. Springer: Heidelberg.

Price K. 1996. Differential evolution: a fast and simple numerical optimizer. In *Proceedings NAFIPS’96*; 524–527, DOI: 10.1109/NAFIPS.1996.534790.

Ryan SE, Porth L, Troendle CA. 2002. Defining phases of bedload transport using piecewise regression. *Earth Surface Processes and Landforms* **27**: 971–990.

Ryan SE, Porth LS. 2007. A tutorial on the piecewise regression approach applied to bedload transport data. *Technical Report General Technical Report RMRS-GTR-189*, United States Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Seim E, Sæther B. 1983. On rethinking allometry: which regression model to use? *Journal of Theoretical Biology* **104**: 161–168.

Smith RJ. 2009. Use and misuse of reduced major axis for line-fitting. *American Journal of Physical Anthropology* **140**: 476–486.

Sokal RR, Rohlf FJ. 2001. *Biometry: The Principles and Practice of Statistics in Biological Research*, (3rd edn). W. H. Freeman and Company: New York.

Stewart J. 2007. *Single Variable Essential Calculus: Early Transcendentals*. Brooks/Cole: Belmont, California.

Storn R, Price K. 1997a. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization* **11**: 341–359.

Storn R, Price K. 1997b. Differential evolution: a simple evolution strategy for fast optimization. *Dr. Dobb’s Journal* **22**: 18–24.

Vieth E. 1989. Fitting piecewise linear regression functions to biological responses. *Journal of Applied Physiology* **67**: 390–396.

Vesterstrøm J, Thomsen R. 2004. A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems. In *Congress on Evolutionary Computation, CEC2004*, Vol. 2. IEEE; L980–L987, DOI: 10.1109/CEC.2004.133tt39.

Walker TJ, Collins N. 2010. New world thermometer crickets: the *Oecanthus rileyi* species group and a new species from North America. *Journal of Orthopera Research* **19**: 371–376. (See supplementary material available at http://www.bioone.org/doi/suppl/10.1665/034.019.0227.).

Zielinski K, Laur R. 2008. Stopping criteria for differential evolution in constrained single-objective optimization. In *Advances in Differential Evolution*, Vol. 143, Chakraborty U (ed.), Studies in Computational Intelligence. Springer-Verlag: Berlin; 111–138.

APPENDIX A. ALGORITHM DESCRIBED IN SECTION 2

To fit the best discontinuous two-phase linear model to data use this algorithm. To extend the algorithm to fit a continuous piecewise linear model, use DE in step 4 to approximate the best piecewise model. The algorithm otherwise proceeds accordingly. Changes for OLS are in parentheses.

1. `Sort` the data pairs, $(x_1, y_1), \ldots, (x_N, y_N)$ by increasing x-value.
2. `Fit` the best RMA (or OLS) regression line and let ϵ_0 be the error in this model as defined by (2) (or (1)).
3. `Obtain` $\left(x_i^*, y_i^*\right)$ as the points on the fitted line where the altitude of the minimized triangles intersects the fitted line (or points on the fitted line with same x-coordinates x_i).

```
4. Compute $v_i = (x_i, y_i) - \left(x_i^*, y_i^*\right)$.
5. Define counter variable $i$.
6. Loop over $i$ from 3 to $n - 3$ (at least three data points are needed to define RMA or OLS line).
    a. Fit the best RMA (or OLS) regression line to $\{(x_1, y_1), \ldots, (x_i, y_i)\}$ and define $\epsilon_{1i}$ as the error in this model.
    b. Fit the best RMA ( or OLS) regression line to $\{(x_{i+1}, y_{i+1}), \ldots, (x_N, y_N)\}$ and define $\epsilon_{2i}$ as the error in this model.
    c. Compute $\epsilon_i = \epsilon_{1i} + \epsilon_{2i}$.
7. Identify the best change-point model as the one which minimizes $\epsilon_i$, set the error of this best model as $\epsilon$.
8. Compute $\Delta\epsilon = \epsilon_0 - \epsilon$.
    Test null hypothesis of no change point:

9. Set $p = 0$. $p$ will be the empirical $p$-value.
10. Bootstrap the residuals: create $S$ replicant data sets by taking the points $\left(x_i^*, y_i^*\right)$ and adding to each such point one of the
    vectors $v_i$ computed in step 4, chosen with replacement.
11. Repeat steps 4, 5, and 6; set $\Delta\epsilon_j$ to be the difference between the best simple linear model for the replicant data set $j$ and the best
    model with a change point for the replicant data set $j$, $j = 1, \ldots, S$.
12. Compute the empirical $p$-value as the proportion of $(1 + \rho)\Delta\epsilon_j > \Delta\epsilon,\ j = 1, \ldots, S$. $\rho \geqslant 0$ is a tuning parameter.
```