Model II regression – User's guide

Pierre Legendre

Département de sciences biologiques, Université de Montréal,
C.P. 6128, succursale Centre-ville, Montréal, Québec H3C 3J7, Canada
E-mail: Pierre.Legendre@umontreal.ca

September 2001

This program computes model II simple linear regression using the following methods: major axis (MA), standard major axis (SMA), ordinary least squares (OLS), and ranged major axis (RMA). Information about these methods is available, for instance, in section 10.3.2 of Legendre and Legendre (1998) and in sections 14.13 and 15.7 of Sokal and Rohlf (1995)*. Parametric 95% confidence intervals are computed for the slope and intercept parameters. A permutation test is available to determine the significance of the slopes of MA, OLS and RMA and also for the correlation coefficient.

Bartlett's three-group model II regression method, described by the above-mentioned authors, is not computed by the program because it suffers several drawbacks. Its main handicap is that the regression lines are not the same depending on whether the grouping (into three groups) is made based on x or y. The regression line is not guaranteed to pass through the centroid of the scatter of points and the slope estimator is not symmetric, i.e. the slope of the regression $y = f(x)$ is not the reciprocal of the slope of the regression $x = f(y)$.

Model II regression should be used when the two variables in the regression equation are random, i.e. not controlled by the researcher. Model I regression using least squares underestimates the slope of the linear relationship between the variables when they both contain error; see example 4 below. Detailed recommendations follow.

Recommendations on the use of model II regression methods

Considering the results of simulation studies, Legendre and Legendre (1998) offer the following recommendations to ecologists who have to estimate the parameters of the functional linear relationships between variables that are random and measured with error (Table 1).

* In Sokal and Rohlf (*Biometry, 2nd edition*, 1981: 551), the numerical result for MA regression for the example data set is wrong. The mistake has been corrected in the 1995 edition.

Table 1 Application of the model II regression methods. The numbers in the left-hand column refer to the corresponding paragraphs in the text.

Par.	Method	Conditions of application	Test possible
1	OLS	Error on y >> error on x	Yes
3	MA	Distribution is bivariate normal Variables are in the same physical units or dimensionless Variance of error about the same for x and y	Yes
4		Distribution is bivariate normal Error variance on each axis proportional to variance of corresponding variable	
4.1	RMA	Check scatter diagram: no outlier	Yes
4.2	SMA	Correlation r is significant	No
5	OLS	Distribution is not bivariate normal Relationship between x and y is linear	Yes
6	OLS	To compute forecasted (fitted) or predicted $\hat{y}$ values (Regression equation and confidence intervals are irrelevant)	Yes
7	MA	To compare observations to model predictions	Yes

1. If the magnitude of the random variation (i.e. the error variance[*]) on the response variable y is much larger (i.e. more than three times) than that on the explanatory variable x, use OLS. Otherwise, proceed as follows.

2. Check whether the data are approximately bivariate normal, either by looking at a scatter diagram or by performing a formal test of significance. If not, attempt transformations to render them bivariate normal. For data that are or can be made to be reasonably bivariate normal, consider recommendations 3 and 4. If not, see recommendation 5.

3. For bivariate normal data, use major axis (MA) regression if both variables are expressed in the same physical units (untransformed variables that were originally measured in the same units) or are dimensionless (e.g. log-transformed variables), if it can reasonably be assumed that the error variances of the variables are approximately equal.

When no information is available on the ratio of the error variances and there is no reason to believe that it would differ from 1, MA may be used provided that the results are interpreted with caution. MA produces unbiased slope estimates and accurate confidence intervals (Jolicoeur, 1990).

[*] Contrary to the sample variance, the error variance on x or y cannot be estimated from the data. An estimate can only be made from knowledge of the way the variables were measured.

MA may also be used with dimensionally heterogeneous variables when the purpose of the analysis is (1) to compare the slopes of the relationships between the same two variables measured under different conditions (e.g. at two or more sampling sites), or (2) to test the hypothesis that the major axis does not significantly differ from a value given by hypothesis (e.g. the relationship $E = b_1 m$ where, according to the famous equation of Einstein, $b_1 = c^2$, c being the speed of light in vacuum).

4. For bivariate normal data, if MA cannot be used because the variables are not expressed in the same physical units or the error variances on the two axes differ, two alternative methods are available to estimate the parameters of the functional linear relationship if it can reasonably be assumed that the error variance on each axis is proportional to the variance of the corresponding variable, i.e. (the error variance of *y* / the sample variance of *y*) ≈ (the error variance of *x* / the sample variance of *x*). This condition is often met with counts (e.g. number of plants or animals) or log-transformed data (McArdle, 1988).

4.1. Ranged major axis regression (RMA) can be used. The method is described below. Prior to RMA, one should check for the presence of outliers, using a scatter diagram of the objects.

4.2. Standard major axis regression (SMA) can be used. One should first test the significance of the correlation coefficient (*r*) to determine if the hypothesis of a relationship is supported. No SMA regression equation should be computed when this condition is not met.

This remains a less-than-ideal solution since SMA slope estimates cannot be tested for significance. Confidence intervals should also be used with caution: simulations have shown that, as the slope departs from ±1, the SMA slope estimate is increasingly biased and the confidence interval includes the true value less and less often. Even when the slope is near ±1 (e.g. example 5), the confidence interval is too narrow if *n* is very small or if the correlation is weak.

5. If the distribution is not bivariate normal and the data cannot be transformed to satisfy that condition (e.g. if the distribution possesses two or several modes), one should wonder whether the slope of a regression line is really an adequate model to describe the functional relationship between the two variables. Since the distribution is not bivariate normal, there seems little reason to apply models such as MA, SMA or RMA, which primarily describe the first principal component of a bivariate normal distribution. So, (1) if the relationship is linear, OLS is recommended to estimate the parameters of the regression line. The significance of the slope should be tested by permutation, however, because the distributional assumptions of parametric testing are not satisfied. (2) If a straight line is not an appropriate model, polynomial or nonlinear regression should be considered.

6. When the purpose of the study is not to estimate the parameters of a functional relationship, but simply to forecast or predict values of *y* for given *x*'s, use OLS in all cases. OLS is the only method that minimizes the squared residuals in *y*. The OLS regression line itself is meaningless. Do not use the standard error and confidence bands, however, unless *x* is known to be free of error (Sokal and Rohlf, 1995: 545, Table 14.3); this warning applies in particular to the 95% confidence intervals computed for OLS by this program.

7. Observations may be compared to the predictions of a statistical or deterministic model (e.g. simulation model) in order to assess the quality of the model. If the model contains random variables measured with error, use MA for the comparison since observations and model predictions should be in the same units.

If the model fits the data well, the slope is expected to be 1 and the intercept 0. A slope that significantly differs from 1 indicates a *difference* between observed and simulated values which is proportional to the observed values. For relative-scale variables, an intercept which significantly differs from 0 suggests the existence of a systematic difference between observations and simulations (Mesplé *et al.*, 1996).

8. With all methods, the confidence intervals are large when n is small; they become smaller as n goes up to about 60, after which they change much more slowly. Model II regression should ideally be applied to data sets containing 60 observations or more. Some of the examples presented below have fewer observations; they are only presented for illustration.

Ranged major axis regression

Ranged major axis regression (RMA) is only described in Legendre and Legendre (1998: 511-512). It is computed as follows:

1. Transform the y and x variables into y' and x', respectively, whose range is 1. Two formulas are available for ranging, depending on the nature of the variables:

- For variables whose variation is expressed relative to an arbitrary zero (interval-scale variables, e.g. temperature in °C), the formula for ranging is:

$$y'_i = \frac{(y_i - y_{min})}{(y_{max} - y_{min})} \quad \text{or} \quad x'_i = \frac{(x_i - x_{min})}{(x_{max} - x_{min})} \qquad \textbf{(1)}$$

- For variables whose variation is expressed relative to a true zero value (ratio-scale or relative-scale variables, e.g. species abundances, or temperature expressed in °K), the recommended formula for ranging assumes a minimum value of 0; eq. 1 reduces to:

$$y'_i = \frac{y_i}{y_{max}} \quad \text{or} \quad x'_i = \frac{x_i}{x_{max}} \qquad \textbf{(2)}$$

2. Compute MA regression between the ranged variables y' and x'. Test the significance of the slope estimate by permutation if needed.

3. Back-transform the estimated slope, as well as its confidence interval limits, to the original units by multiplying them by the ratio of the ranges:

$$b_1 = b'_1 \frac{y_{max} - y_{min}}{x_{max} - x_{min}} \qquad \textbf{(3)}$$

4. Recompute the intercept b_0 and its confidence interval limits, using the original centroid $(\bar{x}, \bar{y})$ of the scatter of points and the estimates of the slope b_1 and its confidence limits:

$$b_0 = \bar{y} - b_1\bar{x} \quad \textbf{(4)}$$

The RMA slope estimator has several desirable properties when the variables *x* and *y* are not expressed in the same units or when the error variances on the two axes differ. (1) The slope estimator scales proportionally to the units of the two variables: the position of the regression line in the scatter of points remains the same irrespective of any linear change of scale of the variables. (2) The estimator is sensitive to the covariance of the variables; this is not the case for SMA. (3) Finally, and contrary to SMA, it is possible to test the hypothesis that an RMA slope estimate is equal to a stated value, in particular 0 or 1. As in MA, the test may be done either by permutation, or by comparing the confidence interval of the slope to the hypothetical value of interest. Thus, whenever MA regression cannot be used because of incommensurable units or because the error variances on the two axes differ, RMA regression can be used. There is no reason, however, to use RMA when MA is justified.

Prior to RMA, one should check for the presence of outliers, using a scatter diagram of the objects. RMA should not be used in the presence of outliers because they cause important changes to the estimates of the ranges of the variables. Outliers that are not aligned fairly well with the dispersion ellipse of the objects may have an undesirable influence on the slope estimate. The identification and treatment of outliers is discussed in Sokal and Rohlf (1995, Section 13.4). Outliers may, in some cases, be eliminated from the data set, or they may be subjected to a *winsorizing* procedure described by these authors.

Input file

Prepare a rectangular data table with rows as objects and columns as variables, without identifiers for the rows or columns. Either the explanatory variable *x* or the response variable *y* can be first; the program asks about the positions of the variables in the file (*x* first or *y* first). The columns are separated either by a tab or by any number of spaces; leading spaces are ignored. Save this data table to an ASCII (text) file.

Output file

The output file contains the following results:

1. Simple regression equation ($\hat{y} = b_0 + b_1 x$) using major axis regression (MA). The regression line is the major axis of the dispersion ellipse, hence its name. It is also called the first principal component of the scatter of objects. Details of the computation are given by Sokal and Rohlf (1995, Box 15.6). The program computes the regression equation as well as parametric 95% confidence intervals for the slope and intercept parameters.

2. Same output, using standard major axis regression (SMA). Standard major axis regression is major axis regression computed from standardized variables; the slope estimate is back-transformed to the original variable units. The slope is estimated by the ratio $b_{SMA} = \pm s_y/s_x$. If the two variables have about the same variance, $b_{SMA} \approx \pm 1$.

3. Same output, using ordinary least squares regression (OLS). The coefficients of correlation (r) and determination (r^2) are also computed. With random data, $b_{OLS} \approx 0$.

4. Same output, using ranged major axis regression (RMA). Ranged major axis regression *is* major axis regression (MA) computed from ranged data. Computation details are given in the previous section. RMA is offered as an option by the program because it entails extra questions about ranging the variables; users of the program may not want to decide about this unless they are specifically interested in the results of RMA regression.

5. Permutation tests may be carried out for the slopes of MA, OLS and RMA and for the correlation coefficient *r*. Results of the tests of b_{OLS} and the correlation r are always identical since these two tests are equivalent. The slope of SMA cannot be tested by permutation (Legendre and Legendre, 1998: 511). When the slope of MA or RMA is larger than 1 (or < –1), the permutation test of significance is carried out on the slope $b' = 1/b$ of the regression of x on y which is smaller than 1 (or > –1) (Legendre and Legendre, 1998: 508). The same applies to RMA; in that case, the test involves the slope as estimated from the ranged data, and not the slope value back-transformed to the original units of the variables.

For the slopes of MA, OLS and RMA, the permutation tests are carried out using the slope estimates b themselves as the reference statistics. In OLS simple linear regression, a permutation test of significance based on the r statistic is equivalent to a permutation test based on the pivotal t-statistic associated with b_{OLS} (Legendre and Legendre, 1998: 21). On the other hand, across the permutations, the slope estimate (b_{OLS}) differs from r by a constant (s_y/s_x) since $b_{OLS} = r_{xy}\ s_y/s_x$, so that b_{OLS} and r are equivalent statistics for permutation testing. As a consequence, a permutation test of b_{OLS} is equivalent to a permutation test carried out using the pivotal t-statistic associated with b_{OLS}. This is not the case in multiple linear regression, however, as shown by Anderson and Legendre (1999).

If the objective is simply to assess the relationship between the two variables under study, one can simply compute the correlation coefficient r and test its significance. A parametric test can be used when the assumption of binormality can safely be assumed to hold, or a permutation test when it cannot.

For the intercept of OLS, the confidence interval is computed using the standard formulas found in textbooks of statistics; results are identical to those of standard statistical software. No such formula, providing correct α-coverage, is known for the other three methods. In the program, the confidence intervals for the intercepts of MA, SMA and RMA are computed by projecting the bounds of the confidence intervals of the slopes onto the ordinate; this results in an underestimation of these confidence intervals.

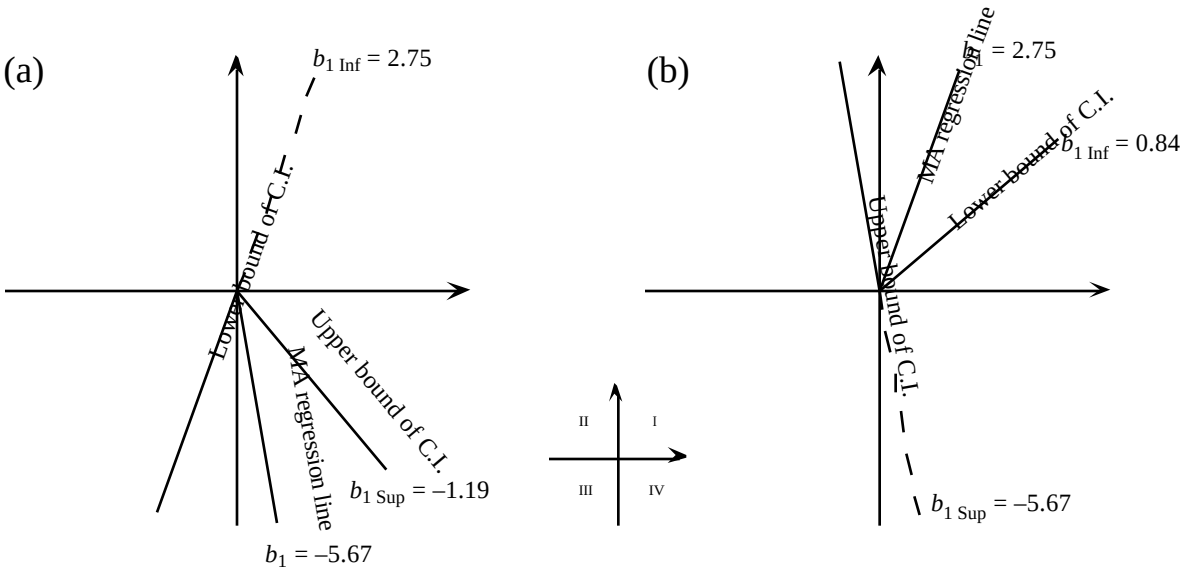

Figure 1 (a) If a MA regression line has the lower bound of its confidence interval (C.I.) in quadrant III, this bound has a positive slope (+2.75 in example). (b) Likewise, if a MA regression line has the upper bound of its confidence interval in quadrant II, this bound has a negative slope (–5.67 in example).

In MA or RMA regression, the bounds of the confidence interval (C.I.) of the slope may, on occasions, lie outside quadrants I and IV of the plane centred on the centroid of the bivariate distribution. When the *lower bound* of the confidence interval corresponds to a line in quadrant III (Fig. 1a), it has a positive slope; the RMA regression line of example 5 provides an example of this phenomenon. Likewise, when the *upper bound* of the confidence interval corresponds to a line in quadrant II (Fig. 1b), it has a negative slope. In other instances, the confidence interval of the slope may occupy all 360° of the plane, which results in it having no bounds. The bounds are then noted 0.00000; see example 5.

In SMA or OLS, confidence interval bounds cannot lie outside quadrants I and IV. In SMA, the regression line always lies at a +45° or –45° angle in the space of the standardized variables; the SMA slope is a back-transformation of ±45° to the units of the original variables. In OLS, the slope is always at an angle closer to zero than the major axis of the dispersion ellipse of the points, i.e. it always underestimates the MA slope in absolute value.

A “secret command” allows users to obtain the list of the reference and permuted values obtained during tests of significance of the slopes of MA and RMA. In answer to the question:

```
Compute ranged major axis regression (RMA)?
(0) no,  (1) yes
```

if one gives –1 as the answer instead of 1, the program writes to a file, called "Permutations", the values of *bma* (slope of MA, which is converted to $1/b_{MA}$ in the program if $b_{MA} > 1$), *brma* (slope of RMA in ranged variable units, which is converted to $1/b_{RMA}$ in the program if $b_{RMA} > 1$) and *brmaC* (slope of RMA converted back to the original units of variables *x* and *y*), as computed in *Subroutine Regression* of the program. The reference (unpermuted) values are found on the first row of that file.

Example 1

Input file

This example compares observations to the values forecasted by a model. A hospital surgical unit wanted to forecast survival of patients undergoing a particular type of liver surgery. Four explanatory variables were measured on patients. The response variable *Y* was survival time, which was $\log_{10}$-transformed. The data are described in detail in Section 8.2 of Neter *et al.* (1996) who also provide the original data sets. The data were divided in two groups of 54 patients. The first group was used to construct forecasting models whereas the second group was reserved for model validation. Several regression models were studied. One of them, which uses variables X_3 = enzyme function test score and X_4 = liver function test score, is used as the basis for the present example. The multiple regression equation is the following:

$$\hat{Y} = 1.388778 + 0.005653\, X_3 + 0.139015\, X_4$$

This equation was applied to the second data set (also 54 patients) to produce forecasted survival times. In the present example, these values are compared to the observed survival times. Fig. 2 shows the scatter diagram with $\log_{10}$(observed survival time) in abscissa and forecasted values in ordinate. The MA regression line is shown with its 95% confidence region. The 45° line, which would correspond to perfect forecasting, is also shown for comparison.

Output file: MA, SMA and OLS equations, 95% C.I., and tests of significance. The RMA method, which is optional, was not computed since MA is the only appropriate method in this example.

```
Model II regression
for situations where variables y and x are both random.
Estimation of the parameters of the functional equation
y = b0 + b1*x
                    © Pierre Legendre, 1994, 1999, 2000
Département de sciences biologiques,  Univ. de Montréal

                ------------------------
```

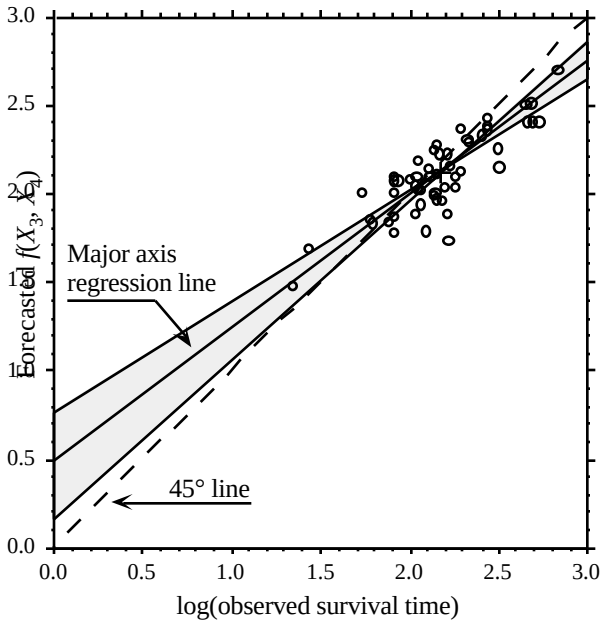

Figure 2 Scatter diagram of the Example 1 data showing the major axis (MA) regression line and its 95% confidence region (grey). The 45° line (dashed) is drawn for reference. The cross indicates the centroid of the bivariate distribution. The MA regression line passes through this centroid.

```
Methods:   major axis (MA),  standard major axis (SMA),
ordinary least-squares (OLS),  ranged major axis (RMA).

Permutation tests: r and slopes of MA, OLS and RMA. The
slope of SMA cannot be tested by permutation (H0:b1=0).

-----------------------------------------------------------

Input data file: EX1_54x2.txt

-------------------------------------------------------------------------

Major axis (MA):
Eigenvalues:   lambda 1 =        0.13324  lambda 2 =        0.01090

b0 =   0.48720    b1 =   0.74921    angle (°) =  36.84093

95% C.I. of slope     = [   0.62166,    0.89456]
95% C.I. of intercept = [   0.17258,    0.76331]    (underestimate)

-------------------------------------------------------------------------

Standard major axis (SMA):

b0 =   0.41155    b1 =   0.78416    angle (°) =  38.10197

95% C.I. of slope     = [   0.67428,    0.91193]
95% C.I. of intercept = [   0.13496,    0.64939]    (underestimate)

C.I. of slope following Jolicoeur & Mosimann (1968), McArdle (1988)
```

```
-----------------------------------------------------------------

Ordinary least-squares (OLS):
          r =   0.83873   coeff. of determination (r^2) =   0.70347

b0 =   0.68530    b1 =   0.65770    angle (°) =  33.33276

95% C.I. of slope     = [   0.53887,    0.77652]
95% C.I. of intercept = [   0.42569,    0.94490]

-----------------------------------------------------------------

b0 = intercept, b1 = slope,  xbar =     2.16466, ybar =     2.10898

Code 999.99999 if the slope is infinite (90° angle).

Code   0.00000 if the limits of the C.I. cannot be computed.
This may happen when the two eigenvalues are too similar;
the C.I. then incorporates all 360° of the plane.


-------------------------------------------
Permutation tests on slopes and correlation
-------------------------------------------

Number of random permutations:   999

-----------------------------------------------
Method    Stat.    LT    EQ    GT   One-tailed p
-----------------------------------------------

MA      0.74921   999     1     0     0.00100

OLS     0.65770   999     1     0     0.00100

Corr    0.83873   999     1     0     0.00100
-----------------------------------------------
```

The interesting aspect of the MA regression equation is that the regression line is not parallel to the 45° line drawn in Fig. 2. The 45° line is not included in the 95% confidence interval of the MA slope, which goes from $\tan^{-1}(0.62166) = 31.87°$ to $\tan^{-1}(0.89456) = 41.81°$. The Figure shows that the forecasting equation overestimates survival below the mean and underestimates it above the mean. Note that the OLS regression line, which is often (erroneously) used by researchers for comparisons of this type, would show an even greater discrepancy (33.3° angle) from the 45° line, compared to the MA regression line (36.8° angle).

Example 2

Input file

The following table presents observations at 20 sites from a study on predator-prey relationships (Hines *et al.* 1998). *y* is the number of bivalves (*Macomona liliana*) larger than 15 mm in size, found in 0.25 m^2 quadrats of sediment; *x* is the number of sediment disturbance pits of a predator, the eagle ray (*Myliobatis tenuicaudatus*), found within circles of a 15 m radius around the bivalve quadrats.

No. predators *x*	No. prey *y*	No. predators *x*	No. prey *y*
2	29	11	53
3	18	5	27
8	29	8	41
4	29	15	58
8	49	7	31
3	33	13	71
8	53	18	59
9	33	6	42
2	20	19	71
4	40	14	54

The variables *x* and *y* are expressed in the same physical units and are estimated with sampling error, and their distribution is approximately bivariate normal. The error variance is not the same for *x* and *y* but, since the data are animal counts, it seems reasonable to assume that the error variance along each axis is proportional to the variance of the corresponding variable. The correlation is significant: $r = 0.86$, $p < 0.001$. RMA and SMA are thus appropriate for this data set; MA and OLS are not. Fig. 3 shows the scatter diagram. The various regression lines are presented to allow their comparison.

Output file: MA, SMA, OLS and RMA regression equations, confidence intervals, and tests of significance (heading removed). That the 95% confidence intervals of the SMA and RMA intercepts do not include 0 may be due to different reasons: (1) the relationship may not be perfectly linear; (2) the C.I. of the intercepts are underestimated; (3) the predators (eagle rays) may not be attracted to sampling locations containing few prey (bivalves).

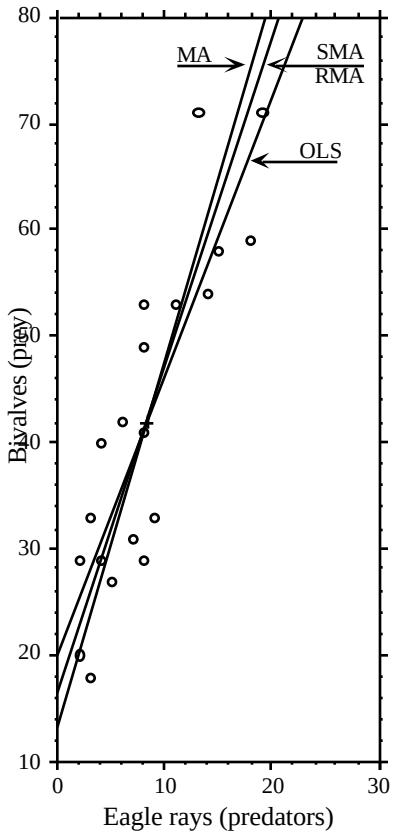

Figure 3 Scatter diagram of the Example 2 data (number of bivalves as a function of the number of eagle rays) showing the major axis (MA), standard major axis (SMA), ordinary least-squares (OLS) and ranged major axis (RMA regression lines. SMA and RMA are the appropriate regression lines in this example. The cross indicates the centroid of the bivariate distribution. The four regression lines pass through this centroid.

```
-----------------------------------------------------

Input data file: EX2_20x2.txt

-----------------------------------------------------------------

Major axis (MA):
Eigenvalues:   lambda 1 =      269.82124  lambda 2 =        6.41823

b0 =  13.05968    b1 =   3.46591    angle (°) =  73.90584

95% C.I. of slope     = [   2.66310,    4.86857]
95% C.I. of intercept = [   1.34742,   19.76310]    (underestimate)

-----------------------------------------------------------------

Standard major axis (SMA):

b0 =  16.45205    b1 =   3.05963    angle (°) =  71.90073

95% C.I. of slope     = [   2.38281,    3.92871]
95% C.I. of intercept = [   9.19529,   22.10353]    (underestimate)

C.I. of slope following Jolicoeur & Mosimann (1968), McArdle (1988)
```

```
------------------------------------------------------------------

Ordinary least-squares (OLS):
          r =   0.86008   coeff. of determination (r^2) =   0.73974

b0 =  20.02675    b1 =   2.63153    angle (°) =  69.19283

95% C.I. of slope     = [   1.85858,    3.40448]
95% C.I. of intercept = [  12.49099,   27.56251]

------------------------------------------------------------------

Ranged major axis (RMA):
ymin =   0.00000 ymax =  71.00000 xmin =   0.00000 xmax =  19.00000
Eigenvalues:   lambda 1 =        0.11509  lambda 2 =        0.00827

b0 =  17.25651    b1 =   2.96329    angle (°) =  71.35239

95% C.I. of slope     = [   2.17426,    3.95653]
95% C.I. of intercept = [   8.96300,   23.84493]    (underestimate)

------------------------------------------------------------------

b0 = intercept, b1 = slope,  xbar =     8.35000, ybar =    42.00000

Code 999.99999 if the slope is infinite (90° angle).

Code   0.00000 if the limits of the C.I. cannot be computed.
This may happen when the two eigenvalues are too similar;
the C.I. then incorporates all 360° of the plane.


-------------------------------------------
Permutation tests on slopes and correlation
-------------------------------------------

Number of random permutations:   999

-----------------------------------------------
Method    Stat.    LT    EQ    GT   One-tailed p                [1]
-----------------------------------------------

MA      3.46591   999     1     0     0.00100

OLS     2.63153   999     1     0     0.00100

Corr    0.86008   999     1     0     0.00100

RMA     2.96329   999     1     0     0.00100
-----------------------------------------------
```

[1] In this table of the output file, the rows correspond, respectively, to the MA, OLS and RMA slopes and to the coefficient of correlation *r* (‘Corr’). ‘Stat.’ is the value of the statistic being tested for significance. As explained in the “Output file” section, the statistic actually used by the program for the test of the MA slope, in this example, is the inverse of the b_{MA}

slope estimate (1/3.46591 = 0.28852) because the reference value of the statistic in this permutation test must not exceed 1.

One-tailed probabilities ('One-tailed p') are computed in the direction of the sign of the coefficient. For a one-tailed test in the upper tail (i.e. for a coefficient with a positive sign), p = (EQ + GT)/(Number of permutations + 1). For a test in the lower tail (i.e. for a coefficient with a negative sign), p = (LT + EQ)/(Number of permutations + 1), where

- LT is the number of values under permutation that are smaller than the reference value;
- EQ is the number of values under permutation that are equal to the reference value of the statistic, plus 1 for the reference value itself;
- GT is the number of values under permutation that are greater than the reference value.

Example 3

Input file

The following table presents data used by Sokal and Rohlf (1995, Box 14.12) to illustrate model II regression analysis. They concern the mass (x) of unspawned females of a California fish, the cabezon (*Scorpaenichthys marmoratus*), and the number of eggs they subsequently produced (y). One may be interested to estimate the functional equation relating the number of eggs to the mass of females before spawning. The physical units of the variables are as in the table published by Sokal and Rohlf (1995: 546).

Mass (×100 g) x	No. eggs (×1000) y
14	61
17	37
24	65
25	69
27	54
33	93
34	87
37	89
40	100
41	90
42	97

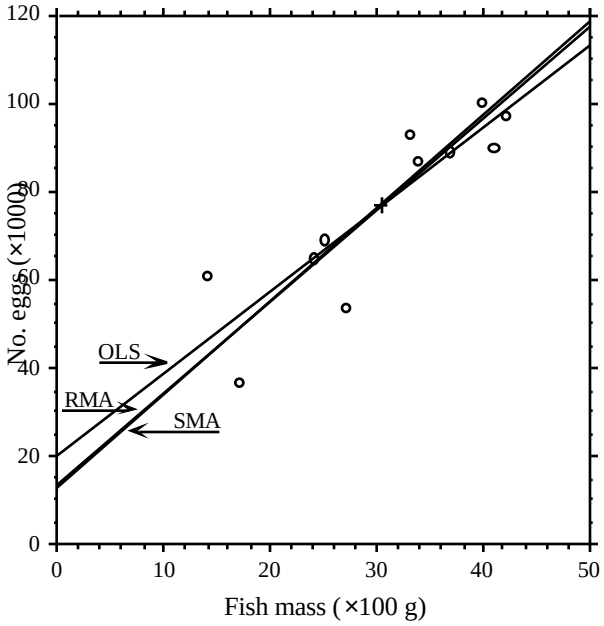

Figure 4 Scatter diagram of the Example 3 data (number of eggs produced as a function of the mass of unspawned females) with the ranged major axis (RMA), standard major axis (SMA) and ordinary least-squares (OLS) regression lines. RMA and SMA are the appropriate regression lines in this example. The cross indicates the centroid of the bivariate distribution. The three regression lines pass through this centroid.

Since the variables are in different physical units and are estimated with error, and their distribution is approximately bivariate normal, RMA and SMA are appropriate for this example; MA is inappropriate. The OLS regression line is meaningless; in model II regression, OLS should only be used for forecasting or prediction. It is plotted in Fig. 4 only to allow comparison.

The RMA and SMA regression lines are nearly indistinguishable in this example. The slope of RMA can be tested for significance (H_0: $b_{RMA} = 0$), however, whereas the SMA slope cannot. The 95% confidence intervals of the intercepts of RMA and SMA, although underestimated, include the value 0, as expected if a linear model applies to the data: a female with a mass of 0 is expected to produce no egg.

Another interesting property of RMA and SMA is that their estimates of slope and intercept change proportionally to changes in the units of measurement. One can easily verify that by changing the decimal places in the Example 2 data file and recomputing the regression equations. RMA and SMA share this property with OLS. MA regression does not have this property; this is why it should only be used with variables that are in the same physical units, as those of Example 1.

Output file: MA, SMA, OLS and RMA equations, 95% C.I., and tests of significance (heading removed).

```
-----------------------------------------------------

Input data file: EX3_11x2.txt

-----------------------------------------------------------------

Major axis (MA):
Eigenvalues:   lambda 1 =      494.63400  lambda 2 =       17.49327

b0 =   6.65663    b1 =   2.30173    angle (°) =  66.51716

95% C.I. of slope     = [   1.60430,    3.72440]
95% C.I. of intercept = [ -36.54082,   27.83295]    (underestimate)

-----------------------------------------------------------------

Standard major axis (SMA):

b0 =  12.19378    b1 =   2.11937    angle (°) =  64.74023

95% C.I. of slope     = [   1.49672,    3.00104]
95% C.I. of intercept = [ -14.57693,   31.09957]    (underestimate)

C.I. of slope following Jolicoeur & Mosimann (1968), McArdle (1988)
-----------------------------------------------------------------

Ordinary least-squares (OLS):
          r =   0.88232   coeff. of determination (r^2) =   0.77849

b0 =  19.76682    b1 =   1.86996    angle (°) =  61.86337

95% C.I. of slope     = [   1.11780,    2.62211]
95% C.I. of intercept = [  -4.09838,   43.63201]

-----------------------------------------------------------------

Ranged major axis (RMA):
ymin =   0.00000 ymax = 100.00000 xmin =   0.00000 xmax =  42.00000
Eigenvalues:   lambda 1 =        0.08926  lambda 2 =        0.00550

b0 =  13.17967    b1 =   2.08690    angle (°) =  64.39718

95% C.I. of slope     = [   1.35993,    3.12944]
95% C.I. of intercept = [ -18.47574,   35.25317]    (underestimate)

-----------------------------------------------------------------

b0 = intercept, b1 = slope,  xbar =    30.36364, ybar =    76.54545

Code 999.99999 if the slope is infinite (90° angle).

Code   0.00000 if the limits of the C.I. cannot be computed.
This may happen when the two eigenvalues are too similar;
the C.I. then incorporates all 360° of the plane.
```

```
------------------------------------------
Permutation tests on slopes and correlation
------------------------------------------

Number of random permutations:   999

-----------------------------------------------
Method    Stat.    LT    EQ    GT   One-tailed p
-----------------------------------------------

MA      2.30173   998     1     1     0.00200

OLS     1.86996   998     1     1     0.00200

Corr    0.88232   998     1     1     0.00200

RMA     2.08690   998     1     1     0.00200
-----------------------------------------------
```

Example 4

Input file

Mesplé *et al.* (1996) generated a variable X containing 100 values drawn at random from a uniform distribution in the interval [0, 10]. They then generated two other variables, N_1 and N_2, containing values drawn at random from a normal distribution $N(0, 1)$. These variables were combined to create two new variables $x = (X + N_1)$ and $y = (X + N_2)$. The relationship constructed in this way between x and y is perfect and should have a slope of 1, despite the fact that there is normal error added independently to x and y.

Output file

The various model II regression methods were applied to this data set with the following results (heading removed):

```
-----------------------------------------------------

Input data file: E4_100x2.txt

-----------------------------------------------------------------

Major axis (MA):
Eigenvalues:   lambda 1 =       17.56697  lambda 2 =        0.95341

b0 =   0.29059    b1 =   0.96029    angle (°) =  43.83962

95% C.I. of slope     = [   0.86957,    1.06006]
95% C.I. of intercept = [  -0.21218,    0.74777]    (underestimate)

-----------------------------------------------------------------

Standard major axis (SMA):

b0 =   0.27034    b1 =   0.96431    angle (°) =  43.95915

95% C.I. of slope     = [   0.88261,    1.05358]
95% C.I. of intercept = [  -0.17951,    0.68207]    (underestimate)

C.I. of slope following Jolicoeur & Mosimann (1968), McArdle (1988)
-----------------------------------------------------------------

Ordinary least-squares (OLS):
          r =   0.89690   coeff. of determination (r^2) =   0.80443

b0 =   0.77135    b1 =   0.86489    angle (°) =  40.85618

95% C.I. of slope     = [   0.77940,    0.95038]
95% C.I. of intercept = [   0.26636,    1.27633]

-----------------------------------------------------------------

Ranged major axis (RMA):
ymin =  -1.51400 ymax =  10.58600 xmin =  -0.81100 xmax =  10.78200
Eigenvalues:   lambda 1 =        0.12558  lambda 2 =        0.00678

b0 =  -0.46075    b1 =   0.95560    angle (°) =  43.69937

95% C.I. of slope     = [   0.86511,    1.05464]
95% C.I. of intercept = [  -0.18483,    0.77021]    (underestimate)

-----------------------------------------------------------------

b0 = intercept, b1 = slope,  xbar =     5.03918, ybar =     5.12968

Code 999.99999 if the slope is infinite (90° angle).

Code   0.00000 if the limits of the C.I. cannot be computed.
This may happen when the two eigenvalues are too similar;
the C.I. then incorporates all 360° of the plane.
```

```
-------------------------------------------
 Permutation tests on slopes and correlation
 -------------------------------------------

 Number of random permutations:   999

 -----------------------------------------------
 Method    Stat.    LT    EQ    GT   One-tailed p
 -----------------------------------------------

 MA      0.96029   999     1     0     0.00100

 OLS     0.86489   999     1     0     0.00100

 Corr    0.89690   999     1     0     0.00100

 RMA     0.95560   999     1     0     0.00100
 -----------------------------------------------
```

The noticeable aspect is that with OLS regression, the confidence interval of the slope does not include the value 1 and the confidence interval of the intercept does not include the value 0. The OLS slope underestimates the real slope of the bivariate functional relationship, which is 1 by construct in this example. This illustrates the fact that OLS, considered as model I regression method, is inadequate to estimate the slope of the functional relationship between these variables. As a model II regression method, OLS would only be appropriate to predict the values $\hat{y}$ from x (point 6 in Table 1).

With all the other model II regression methods, the confidence intervals of the slopes include the value 1 and the confidence intervals of the intercepts include the value 0, as expected for this data set because of the way the data were generated.

Example 5

Input file

Two vectors of 100 random data drawn from a normal distribution $N(0, 1)$ were generated. One expects to find a null correlation with this type of data which were submitted to the model II regression program.

Output file: MA, SMA, OLS and RMA equations, 95% C.I., and tests of significance (heading removed). Fig. 5 shows the scatter diagram. The various regression lines are presented to allow their comparison.

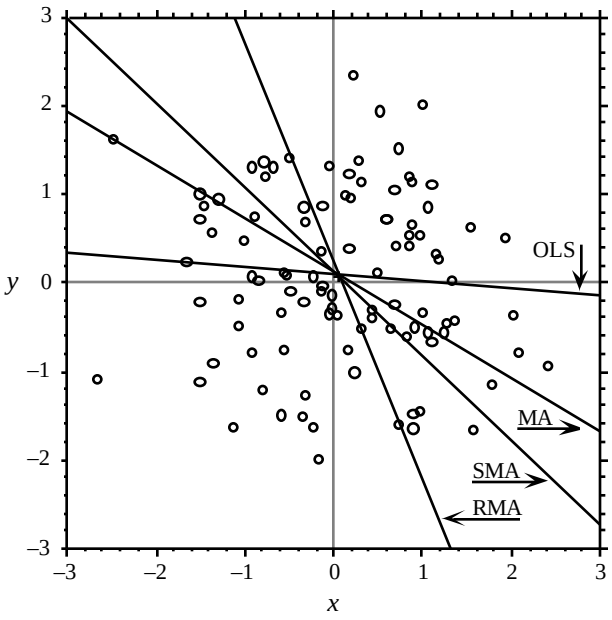

Figure 5 Scatter diagram of the Example 5 data (random numbers) showing the major axis (MA), standard major axis (SMA), ordinary least-squares (OLS) and ranged major axis (RMA regression lines. The correlation coefficient is not significantly different from zero. The cross indicates the centroid of the bivariate distribution. The four regression lines pass through this centroid.

```
------------------------------------------------------------

Input data file: E5_100x2.txt

------------------------------------------------------------------------

Major axis (MA):
Eigenvalues:   lambda 1 =        1.07132  lambda 2 =        0.88571

b0 =   0.11293    b1 =  -0.60005    angle (°) = -30.96569

95% C.I. of slope     = [   0.00000,    0.00000]
95% C.I. of intercept = [   0.00000,    0.00000]    (underestimate)

------------------------------------------------------------------------

Standard major axis (SMA):

b0 =   0.14184    b1 =  -0.95633    angle (°) = -43.72118

95% C.I. of slope     = [  -1.16625,   -0.78419]
95% C.I. of intercept = [   0.12788,    0.15888]    (underestimate)

C.I. of slope following Jolicoeur & Mosimann (1968), McArdle (1988)
------------------------------------------------------------------------

Ordinary least-squares (OLS):
          r =  -0.08377   coeff. of determination (r^2) =   0.00702

b0 =   0.07074    b1 =  -0.08011    angle (°) =  -4.58017

95% C.I. of slope     = [  -0.27114,    0.11092]
95% C.I. of intercept = [  -0.12205,    0.26354]
```

```
-----------------------------------------------------------------

Ranged major axis (RMA):
ymin =  -1.98676 ymax =   2.35266 xmin =  -2.66496 xmax =   2.39702
Eigenvalues:   lambda 1 =        0.05091  lambda 2 =        0.03863

b0 =   0.26978    b1 =  -2.53297    angle (°) = -68.45619

95% C.I. of slope     = [   1.63300,   -0.39805]
95% C.I. of intercept = [   0.09654,   -0.06827]    (underestimate)

-----------------------------------------------------------------

b0 = intercept, b1 = slope,  xbar =     0.08114, ybar =     0.06424

Code 999.99999 if the slope is infinite (90° angle).

Code   0.00000 if the limits of the C.I. cannot be computed.
This may happen when the two eigenvalues are too similar;
the C.I. then incorporates all 360° of the plane.


-------------------------------------------
Permutation tests on slopes and correlation
-------------------------------------------

Number of random permutations:   999

-----------------------------------------------
Method    Stat.    LT    EQ    GT   One-tailed p
-----------------------------------------------

MA     -0.60005   215     1   784     0.21600

OLS    -0.08011   215     1   784     0.21600

Corr   -0.08377   215     1   784     0.21600

RMA    -2.53297   284     1   715     0.28500
-----------------------------------------------
```

Neither the correlation nor any of the regression coefficients are significant; this is as expected from the way the data were generated. Note that the slope estimates differ widely among methods. The MA slope is $b_{MA} = –0.60005$ but its confidence interval, noted [0.00000, 0.00000], covers all 360° of the plane, as stated in the comment underneath the regression table. The RMA slope estimate is $b_{RMA} = –2.53297$. OLS, which should only be used to predict the values $\hat{y}$ from x (point 6 in Table 1), tends to produce slopes near zero for random data: $b_{OLS} = –0.08011$.

Since the correlation is not significant, SMA should not have been computed. This method tends to produce slopes near 1; with the present example, the slope is indeed near 1 ($b_{SMA} = –0.95633$) since the standard deviations of the two variables are nearly equal

(s_x = 1.01103, s_y = 0.96688). This example shows that RMA does not necessarily produce results that are similar to SMA.

The confidence intervals of the slope and intercept of RMA provide an example of the phenomenon of inversion of the confidence limits described in Fig. 1.

Program distribution

A computer program written by P. Legendre is available from our base WWWeb site. Distribution includes the FORTRAN source code, user's manuals, sample files and different versions of the executable program. Versions for MacOS (68k or PowerPC) and 32-bit DOS (suitable for DOS sessions under Windows 95/98/NT) are provided. WWWeb address: <http://www.fas.umontreal.ca/biol/legendre/>.

References

Anderson, M. J. and P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation* **62**: 271-303.

Hines, A. H., R. B. Whitlatch, S. F. Thrush, J. E. Hewitt, V. J. Cummings, P. K. Dayton and P. Legendre. 1997. Nonlinear foraging response of a large marine predator to benthic prey: eagle ray pits and bivalves in a New Zealand sandflat. *Journal of Experimental Marine Biology and Ecology* **216**: 191-210.

Jolicoeur, P. 1990. Bivariate allometry: interval estimation of the slopes of the ordinary and standardized normal major axes and structural relationship. *Journal of Theoretical Biology* **144**: 275-285.

Jolicoeur, P. and J. E. Mosimann. 1968. Intervalles de confiance pour la pente de l'axe majeur d'une distribution normale bidimensionnelle. *Biométrie-Praximétrie* **9**: 121-140.

Legendre, P. and L. Legendre. 1998. *Numerical ecology. 2nd English edition.* Elsevier Science BV, Amsterdam.

McArdle, B. 1988. The structural relationship: regression in biology. *Canadian Journal of Zoology* **66**: 2329-2339.

Mesplé, F., M. Troussellier, C. Casellas and P. Legendre. 1996. Evaluation of simple statistical criteria to qualify a simulation. *Ecological Modelling* **88**: 9-18.

Neter, J., M. H. Kutner, C. J. Nachtsheim and W. Wasserman. 1996. *Applied liear statistical models. 4th Edition.* Richard D. Irwin Inc., Chicago.

Sokal, R. R. and F. J. Rohlf. 1995. *Biometry – The principles and practice of statistics in biological research. 3rd edition.* W. H. Freeman, New York.

Reference to the program

Users of the model II regression program can refer to it through the present user's manual:

Legendre, P. 2001. *Model II regression – User's guide.* Département de sciences biologiques, Université de Montréal. 23 pp.
Available from the WWWeb site <http://www.fas.umontreal.ca/biol/legendre/>.