Determining Functional Relations in Multivariate Oceanographic Systems: Model II Multiple Linear Regression

SCOTT J. RICHTER

Department of Mathematics and Statistics, University of North Carolina at Greensboro, Greensboro, North Carolina

ROBERT H. STAVN

Oceanography Division, Naval Research Laboratory, Stennis Space Center, Mississippi

(Manuscript received 1 October 2013, in final form 3 April 2014)

ABSTRACT

A method for estimating multivariate functional relationships between sets of measured oceanographic, meteorological, and other field data is presented. Model II regression is well known for describing functional relationships between two variables. However, there is little accessible guidance for the researcher wishing to apply model II methods to a multivariate system consisting of three or more variables. This paper describes a straightforward method to extend model II regression to the case of three or more variables.

The multiple model II procedure is applied to an analysis of the optical spectral scattering coefficient measured in the coastal ocean. The spectral scattering coefficient is regressed against both suspended mineral particle concentration and suspended organic particle concentration. The regression coefficients from this analysis provide adjusted estimates of the mineral particle scattering cross section and the organic particle scattering cross section. Greater accuracy and efficiency of the coefficients from this analysis, compared to semiempirical coefficients, is demonstrated. Examples of multivariate data are presented that have been analyzed by partitioning the variables into arbitrary bivariate models. However, in a true multivariate system with correlated predictors, such as a coupled biogeochemical cycle, these bivariate analyses yield incorrect coefficient estimates and may result in large unexplained variance. Employing instead a multivariate model II analysis can alleviate these problems and may be a better choice in these situations.

1. Introduction

Much oceanographic field research seeks to establish a functional relation, that is, find the slope and intercept of the line that best fits the bivariate scatter between two field-observed variables. A good example of this is found in Tett et al. (1975), where particulate carbon and particulate phosphorus are compared. Model II regression has often been suggested in this type of situation to estimate the functional relationship between two variables (Ricker 1973; Sokal and Rohlf 1995; Warton et al. 2006). The literature is substantial regarding estimating a linear functional relationship, dating back to at least Pearson (1901), and several different methods have been proposed. A model II analysis was applied by Sverdrup (1916) in analyzing meteorological variables as early as 1916. Ricker (1973) proposed the use of model II regression in fishery studies. Laws and Archie (1981) asserted the advisability of using model II regression for various field studies, such as the investigation of oxygen consumption per body mass, that is, the metabolic rate, by oceanic zooplankton in the study of Ikeda (1970). Laws and Archie (1981) pointed out that studies of metabolic rate per body mass, phytoplankton ratios of carbon to chlorophyll, and morphometric analyses, all of which are determined from regression slopes, lend themselves to more efficient analyses by model II regression. Further studies of the variations of particulate carbon body mass regressed against chlorophyll concentration of phytoplankton as discussed by Banse (1977) are also amenable to model II analysis.

In virtually all of the studies mentioned above, arbitrary bivariate pairings of variables were modeled, even though the situations investigated actually involved three variables. For example, the studies of Ikeda (1970)

Corresponding author address: Robert H. Stavn, Biology Department, University of North Carolina at Greensboro, 312 Eberhart Building, 321 McIver Street, Greensboro, NC 27402-6170.
E-mail: stavnr@uncg.edu; sjricht2@uncg.edu

DOI: 10.1175/JTECH-D-13-00210.1

involved oxygen uptake, body mass, and environmental temperature, and Banse (1977) considered the variations in cellular particulate phosphorus content and other environmental variables in addition to chlorophyll concentration and particulate organic carbon mass. The multivariate nature of oceanographic investigations is becoming increasingly recognized, as in the studies of Fichot et al. (2008), while Gallie and Murtha (1992) applied a multiple regression analysis to the biogeo-optical properties of inland waters that will be instructive to oceanographers. Stavn and Richter (2008) applied model II multiple regression to determine the mass-specific scattering cross sections of suspended matter in the ocean, both inorganic and organic, to allow more accurate determinations of these materials by satellite remote sensing. All of these situations come under the subject of the interaction of physical properties with coupled biogeochemical cycles (Schlesinger et al. 2011).

In this paper a method for determining model II multiple regression estimates is presented. The method is applied to the determination of optimum mass-specific optical scattering cross sections for suspended matter in the coastal ocean. Section 2 contains a brief background on the development of model II regression methods. Section 3 introduces the notation for the bivariate case, applied to estimate the relation between fish mass and egg production. Section 4 extends the method to more than two variables and provides an example for the three-variable case: determining optical scattering cross sections from oceanographic field measurements. Section 5 contains a discussion of theoretical issues of modeling relations and suggests other possible areas of application of model II multiple regression in oceanography.

2. Model II regression methods

The most widely recommended methods for model II regression are the *major axis* (MA) method and the *standardized* (or *reduced*) *major axis* (SMA), also referred to as the *geometric mean* method. These methods are well developed for the case of two variables (e.g., Laws 1997; Sokal and Rohlf 1995; Warton et al. 2006), but not for three or more variables. Thus, the primary goal of this paper is to present a method of estimating the functional relationship between any number of variables. Many authors have discussed aspects of regression for describing functional relationships, often in a broader context (see Anderson 1984; Kendall and Stuart 1977). However, we do not intend for this to be a general discussion of the various issues of estimation arising in those contexts. Instead, our goal is to outline a straightforward procedure for estimating the parameters in a multivariate functional relationship.

Pearson (1901) showed that the best-fitting line to a system of points is in the direction of the major axis of the corresponding ellipse. Kermack and Haldane (1950) later showed that Pearson's method was not invariant to changes in scale of the variables and proposed the use of a ''reduced'' major axis, where the major axis method is applied to standardized variables. Ricker (1973) showed that geometric mean regression is identical to reduced major axis regression but far easier to compute. Jolicoeur (1975), however, argued against the reduced (or ''standard'') major axis method, and Sprent and Dolby (1980) suggested using the line that bisects the minor angle between the two model I (ordinary least squares) regressions as the estimate of the functional relationship. Warton et al. (2006) argue that the term ''standardized major axis'' is best used to describe the major axis method applied to standardized data, and this is the term that will be used in this paper.

With the exception of Pearson (1901), who also considered the three-variable case, all of the previously mentioned references address only the case of relating two variables. Pearson's 1901 paper is considered the foundation of principal component analysis (PCA), and PCA provides a framework for estimating functional relationships that can easily be extended to more than two variables. While the relationship between PCA and model II methods is generally appreciated, rarely are model II estimates presented using the results of a PCA analysis. Rencher (2002) described the relation between PCA and model II estimates for the two-variable case. In section 3, we present the notation for the two-variable case, and in section 4 we derive estimates for models with three or more variables.

3. Model II regression using PCA

a. Two-variable case

The two-variable case is considered here to introduce the notation required for extension to the multivariate case. One advantage of using PCA is that most statistical software packages (e.g., SAS, R) contain procedures to perform PCA, thus allowing model II estimates to be calculated without special programs.

Consider two variables, Y_1 and Y_2, for which it is desired to estimate the line that best fits the bivariate scatter between the two variables. This is accomplished by minimizing the sum of the squared perpendicular distances from each point to the regression line. To determine the equation of the line that best describes the functional relationship, let $\overline{y}_1$ and $\overline{y}_2$ be the sample means, and s_1^2 and s_2^2 the sample variances of Y_1 and Y_2, respectively; and let s_{12} be the sample covariance between Y_1 and Y_2. Let $\mathbf{a}_1' = (a_{11}, a_{12})$ and $\mathbf{a}_2' = (a_{21}, a_{22})$

TABLE 1. Weights (Y_1) of unspawned female cabezon (a California fish, *Scorpaenichthys marmoratus*) and the number of eggs they subsequently produce (Y_2) for 11 fish.

Fish	1	2	3	4	5	6	7	8	9	10	11
Weight*: Y_1	14	17	24	25	27	33	34	37	40	41	42
Eggs**: Y_2	61	37	65	69	54	93	87	89	100	90	97

* To the nearest 100 g.
** In thousands.

be the eigenvectors associated with the eigenvalues, $\boldsymbol{\lambda}' = (\lambda_1, \lambda_2)$, of the sample covariance matrix $\mathbf{S}$, respectively. Then the principal components are the variables

$$z_1 = a_{11}(y_1 - \bar{y}_1) + a_{12}(y_2 - \bar{y}_2)$$

and

$$z_2 = a_{21}(y_1 - \bar{y}_1) + a_{22}(y_2 - \bar{y}_2).$$

Geometrically, the ordered pairs (y_1, y_2) might be characterized by an ellipse that contains the points. If this ellipse is rotated until the major (longer) axis is horizontal, then the new coordinates of the points are (z_1, z_2), with respect to the new horizontal axis, z_1, and the new vertical axis (perpendicular to z_1), z_2. The major axis is the line passing through the point $\bar{\mathbf{y}}' = (\bar{y}_1, \bar{y}_2)$ in the direction determined by the first eigenvector, $\mathbf{a}_1' = (a_{11}, a_{12})$, which has a slope equal to a_{12}/a_{11}. The equation of the major axis can also be found by setting the second principal component, z_2, to zero. Then,

$$z_2 = a_{21}(y_1 - \bar{y}_1) + a_{22}(y_2 - \bar{y}_2) = 0$$

and

$$y_2 = \left[\bar{y}_2 + \left(\frac{a_{21}}{a_{22}}\right)\bar{y}_1\right] + \left(\frac{-a_{21}}{a_{22}}\right)y_1. \qquad (1)$$

Thus, the major axis is the line with slope $-a_{21}/a_{22}$ and intercept $\bar{y}_2 + (a_{21}/a_{22})\bar{y}_1$. The slope given by $a_{12}/a_{11} = -a_{21}/a_{22}$ is the slope associated with the MA method.

The following example of biological oceanographic data provides an illustration of the ideas described above:

Example 1—Sokal and Rohlf (1995) reported weights (Y_1) of unspawned female cabezon (a California marine fish, *Scorpaenichthys marmoratus*) and the number of eggs subsequently produced (Y_2) for 11 fish. The data are given in Table 1.

For the data from Table 1, $\bar{y}_1 = 30.364$, $\bar{y}_2 = 76.545$, $s_1^2 = 93.355$, $s_2^2 = 418.873$, and $s_{12} = 174.382$. Then $\mathbf{a}_1' = (0.398, 0.917)$ and $\mathbf{a}_2' = (0.917, -0.398)$ are the eigenvectors associated with the eigenvalues, $\boldsymbol{\lambda}' = (494.6, 17.5)$, and the major axis has slope $(-0.917)/(-0.398) = 2.304$ and intercept $76.545 + (0.917 \times 30.364)/(-0.398) = 6.586$, and thus the equation is $y_2 = 6.586 + 2.304y_1$, illustrated in Fig. 1.

b. *Using standardized data*

When the variables have substantially different variances, the variable with the larger variance will dominate the eigenvector, in the sense that the weight assigned to that variable will be much larger. This can occur if Y_1 and Y_2 are measured in different units, or when the units are the same but the variation in observations is much larger for one of the variables. In example 1, notice that $s_{Y_1}^2 = 93.4$, while $s_{Y_2}^2 = 418.9$. The resulting slope estimate from the major axis regression is 2.30, which is much closer to the model I estimate obtained by regressing Y_1 on Y_2 (slope = 2.38) than the model 1 estimate obtained by regressing Y_2 on Y_1 (slope = 1.90). While the true slope of the relationship is unknown, this property of MA regression has often been criticized (e.g., Ricker 1973; Sokal and Rohlf 1995). In such a case it is often recommended that the variables be standardized before computing the model II estimates. This is equivalent to performing principal component analysis on the correlation, rather than covariance, matrix. In this case, as before, the equation of the major axis is

$$z_2 = a_{21}'(y_1 - \bar{y}_1)/s_1 + a_{22}'(y_2 - \bar{y}_2)/s_2 = 0$$

and

$$y_2 = \left[\bar{y}_2 + \left(\frac{s_2}{s_1}\right)\left(\frac{a_{21}'}{a_{22}'}\right)\bar{y}_1\right] + \left(\frac{s_2}{s_1}\right)\left(\frac{-a_{21}'}{a_{22}'}\right)y_1, \qquad (2)$$

where a_{21}' and a_{22}' now represent the eigenvector components from the correlation matrix. With standardized data, variables contribute equally to the principal components, and the magnitudes of a_{21}' and a_{22}' will always be the same, and thus the slope estimator reduces to $\pm(s_2/s_1)$. This is the SMA estimator.

If standardized data are used for the Cabezon weight and egg production data in example 1, then the eigenvectors of the correlation matrix are $\mathbf{a}_1 = (0.707, 0.707)$ and $\mathbf{a}_2 = (0.707, -0.707)$. The slope is then $\sqrt{(418.873/93.355)}(-0.707/-0.707) = 2.118$, and the intercept is $76.545 + \sqrt{(418.873/93.355)}(0.707/-0.707) \times 30.364 = 12.227$. Substituting this value into Eq. (2) yields the equation $y_2 = 12.227 + 2.118y_1$, illustrated in Fig. 1. The slope estimate, 2.118, can be shown to be the geometric mean of the two least squares slopes.

4. Extension to more than two variables

a. *PCA with three variables*

The results of the previous section can be extended to the case of three variables, Y_1, Y_2, and Y_3. Geometrically,

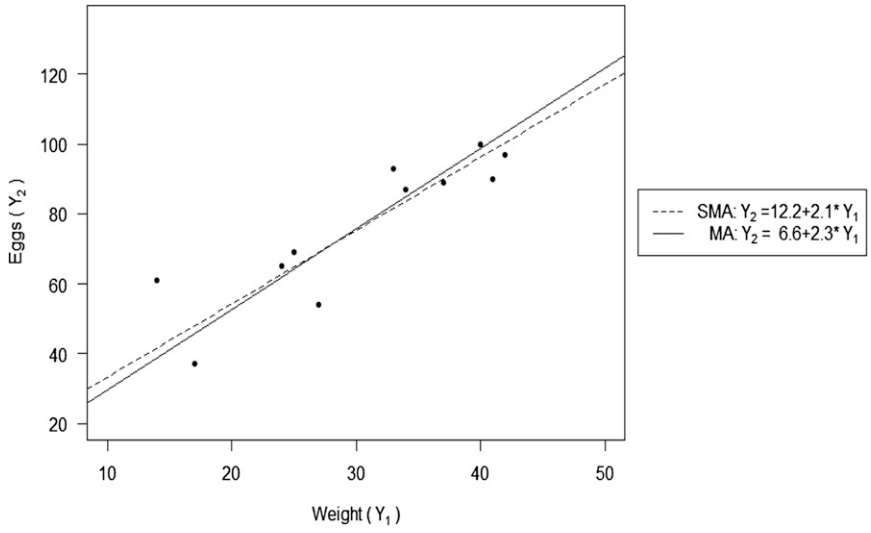

FIG. 1. Scatterplot of Y_2 (number of eggs from each female after spawning) vs Y_1 (weight of unspawned female Cabezon) from Table 1, with MA and SMA fitted lines superimposed. Note the similarity of the scales of the two axes.

the data now form a three-dimensional cloud, and if the variables are jointly correlated, then the cloud will be shaped like an ellipsoid (more or less shaped like an American football). Let $\bar{y}_1, \bar{y}_2$, and $\bar{y}_3$ be the sample means and s_1^2, s_2^2, and s_3^2 the sample variances of Y_1, Y_2 and Y_3, respectively; and let s_{12}, s_{13}, and s_{23} be the sample covariances between Y_1 and Y_2, Y_1 and Y_3, and Y_2 and Y_3, respectively. Let $\mathbf{a}_1' = (a_{11}, a_{12}, a_{13})$, $\mathbf{a}_2' = (a_{21}, a_{22}, a_{23})$, and $\mathbf{a}_3' = (a_{31}, a_{32}, a_{33})$ be the eigenvectors of the sample covariance matrix $\mathbf{S}$. Then the principal components are the variables

$$z_1 = a_{11}(y_1 - \bar{y}_1) + a_{12}(y_2 - \bar{y}_2) + a_{13}(y_3 - \bar{y}_3),$$
$$z_2 = a_{21}(y_1 - \bar{y}_1) + a_{22}(y_2 - \bar{y}_2) + a_{23}(y_3 - \bar{y}_3), \quad \text{and}$$
$$z_3 = a_{31}(y_1 - \bar{y}_1) + a_{32}(y_2 - \bar{y}_2) + a_{33}(y_3 - \bar{y}_3).$$

To derive the coefficients of the function relating Y_3 to Y_1 and Y_2, set $z_3 = 0$ and solve for y_3:

$$a_{31}(y_1 - \bar{y}_1) + a_{32}(y_2 - \bar{y}_2) + a_{33}(y_3 - \bar{y}_3) = 0$$

and

$$y_3 = \left(\frac{a_{31}}{a_{33}}\bar{y}_1 + \frac{a_{32}}{a_{33}}\bar{y}_2 + \bar{y}_3\right) + \left(-\frac{a_{31}}{a_{33}}\right)y_1 + \left(-\frac{a_{32}}{a_{33}}\right)y_2. \tag{3}$$

The following example involving oceanographic field data is used to illustrate the method:

Example 2—Oceanographic field data were collected on the optical scattering coefficient at a wavelength of 555 nm (b_{555}), the particulate inorganic matter concentration (PIM), and the particulate organic matter concentration (POM) of surface samples, from 24 field stations at Mobile Bay, Alabama (see Stavn and Richter 2008). The data are given in Table 2.

For the data from Table 2, $\bar{y}_1 = 6.849$, $\bar{y}_2 = 1.738$, $\bar{y}_3 = 5.724$; $s_1^2 = 20.777$, $s_2^2 = 0.950$, $s_3^2 = 14.696$; and $s_{12} = 2.533$, $s_{13} = 16.471$, $s_{23} = 2.550$. Then $\mathbf{a}_1' = (0.764, 0.105, 0.637)$, $\mathbf{a}_2' = (0.598, -0.485, -0.637)$, and $\mathbf{a}_3' = (-0.242, -0.868, 0.434)$ are the eigenvectors associated with the eigenvalues, $\boldsymbol{\lambda}' = (34.864, 1.175, 0.383)$. The coefficient associated with PIM is $-(-0.242/0.434) = 0.558$, the coefficient associated with POM is $-(-0.868/0.434) = 2.000$, and the intercept is $(-0.242/4.334)(6.849) + (-0.868/0.434)(1.738) = -1.571$, which yields the MA model II equation $b_{555} = -1.571 + 0.558 \times \text{PIM} + 2.000 \times \text{POM}$.

b. *Using standardized data*

The principal components for the standardized data are

$$z_1 = a_{11}'(y_1 - \bar{y}_1)/s_1 + a_{12}'(y_2 - \bar{y}_2)s_2 + a_{13}'(y_3 - \bar{y}_3)s_3,$$
$$z_2 = a_{21}'(y_1 - \bar{y}_1)/s_1 + a_{22}'(y_2 - \bar{y}_2)s_2 + a_{23}'(y_3 - \bar{y}_3)s_3, \quad \text{and}$$
$$z_3 = a_{31}'(y_1 - \bar{y}_1)/s_1 + a_{32}'(y_2 - \bar{y}_2)s_2 + a_{33}'(y_3 - \bar{y}_3)s_3.$$

The equation of the plane determined by the first two principal components can be found by setting $z_3 = 0$:

$$a_{31}'(y_1 - \bar{y}_1)/s_1 + a_{32}'(y_2 - \bar{y}_2)s_2 + a_{33}'(y_3 - \bar{y}_3)s_3 = 0,$$

TABLE 2. Data on the total scattering coefficient at a wavelength of 555 nm [b_{555} (m^{-1})], the concentration of particulate inorganic matter [PIM ($g\,m^{-3}$)], and particulate organic matter [POM ($g\,m^{-3}$)] of surface samples from 24 field stations at Mobile Bay, Alabama.

PIM: Y_1	11.36 6.98 6.89 14.60 12.52 5.40 6.45 1.57 2.15 22.31 4.67 5.01 5.33 5.46 9.98 5.67 6.89 3.21 4.56 6.56 4.63 5.48 3.88 2.80
POM: Y_2	2.36 1.49 1.15 3.00 1.59 2.53 2.21 0.18 0.45 3.28 2.05 0.52 1.94 2.16 2.87 1.81 3.11 3.32 1.69 1.06 1.05 0.69 0.71 0.48
b_{555}: Y_3	8.41 5.85 6.91 11.31 10.03 3.40 5.43 1.09 1.84 17.94 4.85 1.20 4.99 5.05 9.11 8.39 7.57 4.38 3.78 4.91 3.09 3.66 2.41 1.77

and

$$y_3 = [(a'_{31}/a'_{33})(s_3/s_1)\bar{y}_1 + (a'_{32}/a'_{33})(s_3/s_2)\bar{y}_2 + \bar{y}_3] + (-a'_{31}/a'_{33})(s_3/s_1)y_1 + (-a'_{32}/a'_{33})(s_3/s_2)y_2. \quad (4)$$

Thus, the slope coefficients are $(-a'_{31}/a'_{33})(s_3/s_1)$ for Y_1 and $(-a_{32}/a_{33})(s_3/s_2)$ for Y_2. Unlike the two-variable case [Eq. (2)], the ratio of eigenvector elements will usually not equal 1, and thus the slope estimates do not reduce to the ratio of sample standard deviations.

Returning to example 2:

The correlation matrix for Y_1, Y_2, and Y_3 is

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.570 & 0.943 \\ 0.570 & 1.000 & 0.683 \\ 0.943 & 0.683 & 1.000 \end{pmatrix},$$

which has eigenvectors $\mathbf{a}'_1 = (0.594, 0.515, 0.618)$, $\mathbf{a}'_2 = (0.474 - 0.845, -0.249)$, and $\mathbf{a}'_3 = (0.650, 0.145, -0.746)$ associated with the eigenvalues, $\boldsymbol{\lambda}' = (2.476, 0.479, 0.046)$. Then the slope coefficient associated with PIM is $(0.650/-0.746)(3.834/4.558) = -0.733$, the slope coefficient associated with POM is $(0.145/-0.746)(3.834/0.975) = -0.764$, and the intercept is $(0.650/-0.746)(3.834/4.558)(6.849) + (0.145/-0.746)(3.834/0.975)(1.738) + 5.724 = -0.625$, which, when substituted into Eq. (4), yields the SMA model II equation $b_{555} = -0.625 + 0.733 \times \text{PIM} + 0.764 \times \text{POM}$.

These values of b_{555} are compared with values determined in the same way from the Southwest Pass, Mississippi River, a theoretical calculation for suspended coccolithophore plates, and values of a claimed empirical proxy for the slope coefficient in Table 3. The R code (R Core Team 2012) used to perform the calculations in sections 4a and 4b is included in the appendix.

c. *Interpretation of slope coefficients*

As in model I multiple regression, each of the coefficients of y_1 and y_2 can be interpreted as a conditional slope, holding the other variable constant. Geometrically, Eq. (4) represents a plane hovering above the plane determined by y_1 and y_2, as illustrated in Fig. 2. Conditioning on a value of y_2, for example, is equivalent to taking a slice (which is a line) out of the plane and placing it on the y_1y_3 plane, where the value $(-a_{31}/a_{33})$ gives the slope of that line; that is, it gives a description of the linear relationship between y_1 and y_3 for a fixed value of y_2. As is the case in model I regression, the interpretation of the conditional slopes becomes more complicated when the variables are correlated, as it may not be possible to change one variable value while holding the other constant.

d. *Extension to more than three variables*

The results of sections 4a and 4b can be extended to more than three variables. In general, suppose there are $p \geq 2$ observed variables, $Y_1, Y_2, \ldots, Y_p$. The MA equation, using the covariance matrix, becomes

$$y_p = \left[\sum_{i=1}^{p-1} (a_{pi}/a_{pp})\bar{y}_i + \bar{y}_p\right] + \sum_{i=1}^{p-1} (-a_{pi}/a_{pp})y_i \quad (5)$$

and thus the conditional slope of y_k is $(-a_{pi}/a_{pp})$. When using the correlation matrix, the SMA equation becomes

$$y_p = \left[\sum_{i=1}^{p-1} (a'_{pi}/a'_{pp})(s_p/s_i)\bar{y}_i + \bar{y}_p\right] + \sum_{i=1}^{p-1} (-a'_{pi}/a'_{pp})(s_p/s_i)y_i. \quad (6)$$

e. *Adjusted versus unadjusted coefficients*

When considering the analysis of environmental data, collected over various nonconstant conditions of environment, it becomes necessary to factor in other relevant variables in addition to a single variable pair of interest. Therefore, most models for environmental data are actually multivariate in nature. Coefficients adjusted for other variables are easy to obtain using model I regression, but in situations where model II regression is more appropriate, researchers have usually resorted to fitting separate one-predictor models to obtain coefficient estimates (Banse 1977; Ikeda 1970; Steele and Baird 1961). In such a situation, if the explanatory variables are related, then regression applied in turn to variable pairs will result in incorrect estimates of the regression coefficients. Multiple predictor models provide more appropriate estimates than separate one-predictor models, especially when there is substantial dependence between the predictors.

TABLE 3. Mass-specific scattering cross sections at 555 nm compared with mass-specific scattering coefficients at 555 nm (Babin et al. 2003). The asterisk represents the calculation that is from theoretical data at 532 nm reported in Stavn and Richter (2008).

Area or source	$\sigma_{(PIM)}$ ($m^2 g^{-1}$)	$\sigma_{(POM)}$ ($m^2 g^{-1}$)	Source	b^*_{TSS} ($m^2 g^{-1}$)
Mobile Bay	0.73	0.76	Open ocean average	1.00
Southwest Pass	0.57	0.67	Coastal ocean average	0.50
Lith plates of *E. huxleyi**	0.98	—	—	—

Consider the coefficients obtained from Eq. (4) in section 4b, using standardized data. If separate one-predictor model II regression models, using Eq. (2), are fit to the data of Table 2, then the unadjusted slope coefficients for PIM and POM are 0.841 and 3.934, respectively. Recall that the estimates obtained from the SMA method were 0.733 and 0.764, respectively, which are substantially different. This is not surprising, since the variables PIM and POM are moderately correlated ($r = 0.57$). Note also that the same phenomenon occurs using model I regression, where the multiple model I regression model estimates the coefficients to be 0.690 and 0.846 for PIM and POM, respectively, compared to 0.793 and 2.686, respectively, for the separate model I univariate regression models.

f. Confidence intervals for slope coefficients

Once slope estimates are obtained, it might be desirable to provide confidence limits on the true parameter values. Anderson (1963) discussed approximate confidence intervals for eigenvalues under asymptotic multivariate normality. Jolicoeur (1968) proposed a small sample interval for the two-variable case under bivariate normality. Clarke and Van Gorder (2013) derived the density of the ratio of the true regression coefficient to that of the true geometric mean coefficient and use it to construct confidence intervals for the true regression coefficient, also for the two-variable case. Alternatively, nonparametric multivariate bootstrapping may be used to construct confidence limits for the slope coefficients. Advantages of the bootstrap technique are that no assumptions regarding the form of the joint distribution of the variables are necessary, and it can be used to calculate confidence limits using estimates obtained from both standardized and unstandardized data.

Consider the two-variable case discussed in section 3a, and let the true slope coefficient be denoted by θ. To construct the bootstrap distribution, randomly sample, with replacement, n bivariate pairs from the original sample and compute the slope coefficient for each sample. Let the quantity $\hat{\theta}_B = (-a_{21_B}/a_{22_B})$ denote the slope coefficient associated with a particular bootstrap sample. This is repeated for many random samples from the original sample and for the values of $\hat{\theta}_B$ collected to obtain an estimate of the sampling distribution of $\hat{\theta}$. Quantiles of this bootstrap sampling distribution can be used to provide confidence limits for the true coefficient. For $100(1-\alpha)\%$ confidence, the interval is $(\hat{\theta}_{B,\alpha/2}, \hat{\theta}_{B,1-\alpha/2})$. Recall from example 1 in section 3a, the estimated coefficient from the MA method was calculated as $\hat{\theta} = 2.304$. Using 10 000 bivariate bootstrap samples, a 95% confidence interval for the true coefficient is found to be $1.443 \leq \boldsymbol{\theta} \leq 3.163$.

For the multivariate case, the same process can be used. For example, for the three-variable case described in example 2 of section 4a random triplets (PIM, POM, b_{555}) are selected, with replacement, and the slope coefficients for PIM and POM calculated each time. This results in two bootstrap distributions, one for each of the coefficients. Then the appropriate quantiles of the bootstrap sampling distributions provide confidence limits. This process can be extended to any number of variables. In example 2, the estimates using the MA method [Eq. (3)] were 0.558 and 2.000 for PIM and POM, respectively. Again, using 10 000 multivariate bootstrap samples, the 95% confidence intervals are $0.159 \leq \theta_{PIM} \leq 0.754$ for the coefficient of PIM and $0.750 \leq \theta_{POM} \leq 6.382$ for the coefficient of POM.

Note that in example 2, the variability in the observed values of PIM was substantially higher than that of POM (the variance of PIM is almost 24 times as large). The result is a much larger slope estimate for POM compared to PIM, and a great deal of variability in the confidence interval estimates. As was discussed in section 3b, when the variability of the variables is substantially different, calculations on standardized values may yield more useful results. Using the SMA method [Eq. (4)], the slope for PIM is estimated as 0.733, with 95% confidence interval $0.558 \leq \theta_{PIM} \leq 0.825$, while the slope estimate for POM is 0.764, with 95% confidence interval $0.121 \leq \theta_{POM} \leq 1.730$.

5. Discussion

The importance of multivariate models in oceanography and environmental studies is becoming increasingly recognized, and more studies are taking advantage of this point of view. In situations where determining the best functional relation is the goal, the results in section 4 provide straightforward extensions of the MA and

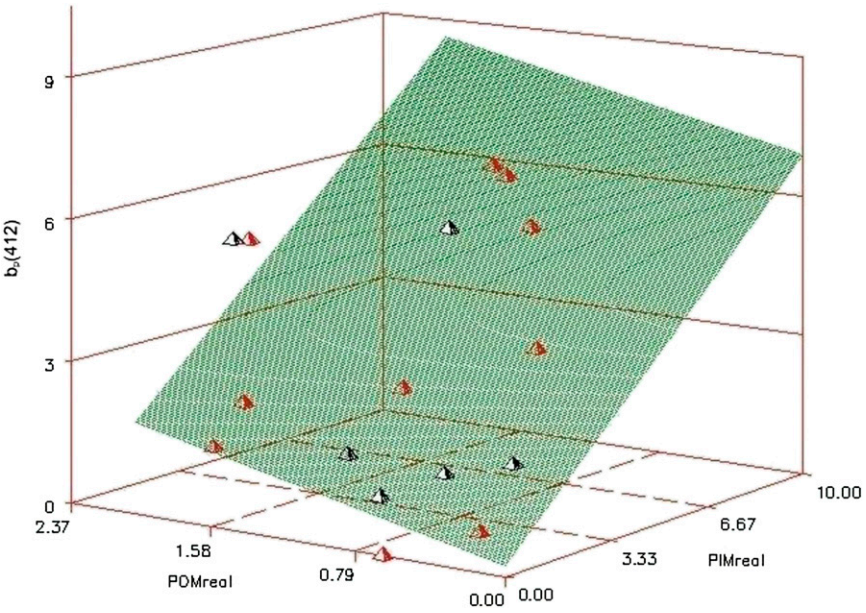

FIG. 2. Three-dimensional plot of suspended mineral (PIMreal; mg L^{-1}) concentration (Y_1), organic matter (POMreal; mg L^{-1}) concentration (Y_2), and optical scattering coefficient (Y_3) at 412 nm [b_p(412); m^{-1}]. Plane represents predicted b_p(412) values from first two principal components of standardized MA analysis. Observed b_p(412) values are black above plane and red below plane.

SMA methods that can be used to find slope coefficients to estimate a functional relation in a multivariate system. There are many examples in the literature where the parameters of a multivariate model have been estimated using estimates from a series of bivariate models. However, this can result in incorrect estimates, since when the variables in the system are correlated, the slope coefficient obtained in a bivariate model is affected by the omitted variables.

There have been attempts to perform analyses of the ocean environment similar to that possible with model II multiple regression by a bivariate approximation to the mass-specific scattering cross section called the mass-specific scattering coefficient, which is equal to b_λ/TSS, the ratio of the total volume scattering coefficient to the total suspended solids (the sum of suspended mineral matter and suspended organic matter) (Babin et al. 2003). There are two problems with this empirical ratio: the volume total scattering coefficient, b_λ, is not partitioned into mineral and organic components, and TSS is not partitioned into mineral and organic components. This empirical ratio has then been recorded for coastal waters and open ocean waters, the average value of each oceanic system being 1.0 and 0.5 $m^2\,g^{-1}$ at 555 nm, respectively, declared as a mass-specific scattering coefficient for open ocean and coastal ocean, respectively (Babin et al. 2003). This assertion has been based on assuming that coastal ocean water contains ''mostly suspended inorganic matter'' and that the open ocean water contains ''mostly suspended organic matter.'' The mass-specific scattering coefficient has then been used in the same manner as the true optical mass-specific scattering cross sections. The determinations of the mass-specific scattering cross sections for the northern Gulf of Mexico are illustrated in Table 3. We can see immediately that the empirical ratio reported in Babin et al. (2003) provides no insight into the optical properties and therefore the efficacy of this bivariate parameter for modeling and prediction of photon budgets, etc., of the northern coastal Gulf of Mexico.

The assertion that the approximation to the mass-specific scattering cross section provided by the coastal average of b_{555}/TSS is 0.5 $m^2\,g^{-1}$ simply does not stand up to scrutiny. Stavn and Richter (2008) used multivariate model II methods to determine the functional relations of the optical volume scattering coefficient of the ocean hydrosol and the concentration of suspended mineral and organic matter in the ocean. The model II multivariate analysis yields the mass-specific optical scattering cross sections of suspended mineral and organic matter. These optical cross sections allow the inversion of the suspended mass of mineral and organic matter by their separate contributions to the remote sensing reflectance signal detected by satellites and

aircraft. Their magnitude and spectral slope are related to the modal size of the suspended matter size distribution and their fractal dimensions, respectively (Stavn 2012). Furthermore, these cross sections allow the partitioning of the hydrosol scattering coefficient into the scattering coefficient due to suspended mineral matter and the scattering coefficient due to suspended organic matter. From these relations it is possible to derive algorithms and models of photon penetration and reflectance, determine a photon budget, and quantitatively account for the components of the remote sensing signal. The values of $\sigma_{(\mathrm{PIM})}$ reported in Stavn and Richter (2008) vary from 0.57 to 0.76 $\mathrm{m^2\,g^{-1}}$ for the northern Gulf of Mexico in a spectral range comparable to the results reported by Babin et al. (2003). Furthermore, the organic values in Table 3 are significantly less than the organic value of 1.0 $\mathrm{m^2\,g^{-1}}$ reported for the empirical ratio, while the Southwest Pass organic value actually approaches the value of the empirical ratio reported for suspended minerals. The closest value of Table 3 to that reported for suspended organic matter (Babin et al. 2003) is the 0.98 $\mathrm{m^2\,g^{-1}}$ calculated for lith plates by Stavn and Richter (2008), except that the lith plates are essentially pure mineral, $CaCO_3$, that would be expected from coccolithophores growing in the open ocean. We see, then, that the empirical ratio b_λ/TSS does not supply unequivocal information about the optical nature nor the type of suspended matter in the coastal ocean.

Model II methods are often recommended when all variables are measured subject to error (Sokol and Rholf 1995). However, as Warton et al. (2006) point out, this is not the most effective justification for using model II methods. Rather, model II methods should be preferred when the goal is to estimate the linear relation that best describes the scatter of the variables. An example of this use was the estimation of scattering coefficients discussed earlier in this section. Alternatively, if the goal is prediction of one variable value based on the values of one or more predictors, then model I regression is usually preferred.

It should also be noted that in addition to potential measurement error, which also may include sampling error, equation error may also exist. If the equation is misspecified, then estimates for both model I and model II methods will generally be biased. Further research is needed to better understand properties of multiple model II estimates.

6. Conclusions

A straightforward procedure for estimating parameters in a model II multiple regression has been presented. The utilization of MA and SMA and where each may be most appropriate was investigated. The procedure was applied to determine the mass-specific scattering cross section of suspended mineral and organic matter in the ocean. Many field situations presently investigated or analyzed as bivariate regressions are actually embedded in a multivariate system. Multivariate estimation is preferred in these situations to obtain more accurate estimates.

Acknowledgments. SJR wishes to acknowledge support, in part, by a University of North Carolina at Greensboro Research Assignment. RHS wishes to acknowledge the valuable discussions on regression from Charles C. Trees, Center for Hydro-Optics and Remote Sensing, San Diego State University, San Diego California. In addition there was the valuable support of a Research Assignment Leave from the University of North Carolina at Greensboro; a National Research Council Research Associateship through NASA, Applied Sciences Directorate, Stennis Space Center, Mississippi; a summer fellowship from the American Society for Engineering Education through the Naval Research Laboratory, Code 7323, Coastal and Marginal Seas Section, Stennis Space Center, Mississippi; early support from ONR Grant N00014-97-0812; and the Naval Research Lab Project to Richard W. Gould Jr., ''Predicting Coastal Bio-Optical Response to Atmospheric/Oceanographic Forcing,'' Stennis Space Center, Mississippi.

APPENDIX

R Code

```
#Analysis of Table 2 data
#Enter data vectors. (Each variable entered as a separate vector)

PIM=c(11.36055,6.98143,6.89490,14.60099,12.52414,
  5.40229,6.45010,1.56994,2.15346,22.30509,4.67338,
  5.00835,5.32663,5.45830,9.98043,5.67296,6.88813,
  3.20948,4.56163,6.55727,4.62544,5.48337,3.87877,
  2.80227)
POM=c(2.35529,1.48723,1.15241,3.00136,1.59180,
  2.53472,2.20666,0.17953,0.45311,3.28352,2.05125,
  0.52108,1.94375,2.16183,2.86684,1.80909,3.10551,
  3.32273,1.69203,1.06135,1.05097,0.69146,0.71080,
  0.48471)
b555 = c(8.40735,5.85045,6.90855,11.30800,10.02730,
  3.39939,5.43069,1.08864,1.84174,17.94280,4.85276,
  1.19745,4.99449,5.05150,9.11405,8.39094,7.57381,
  4.38397,3.77923,4.91364,3.09047,3.66245,2.40817,
  1.76638)
```

```
#Combine vectors into single data matrix

combined=cbind(PIM,POM,b555)

#Compute sample size

n=nrow(combined)

#Compute means and standard deviations and display results

ybar_PIM=mean(PIM)
ybar_POM=mean(POM)
ybar_b555 = mean(b555)
sd_PIM=sd(PIM)
sd_POM=sd(POM)
sd_b555 = sd(b555)
ybar_PIM
ybar_POM
ybar_b555
sd_PIM
sd_POM
sd_b555

#Compute covariance (S) and correlation (R) matrices and display results

S=cov(combined)
R=cor(combined)
S
R

#Calculate eigenvalues and eigenvectors and display results

PCA_S=eigen(S)
PCA_R=eigen(R)
PCA_S
PCA_R

#Compute intercept and slope estimates presented in section 4a, and display results

intercept_S=PCA_S$vectors[1,3]/PCA_S$vectors
  [3,3]*mean(PIM)+PCA_S$vectors[2,3]/PCA_S
  $vectors[3,3]*mean(POM)+mean(b555)
slope_S.PIM=-PCA_S$vectors[1,3]/PCA_S$vectors
  [3,3]
slope_S.POM=-PCA_S$vectors[2,3]/PCA_S$vectors
  [3,3]
intercept_S
slope_S.PIM
slope_S.POM

#Compute intercept and slope estimates presented in section 4b, and display results

intercept_R=PCA_R$vectors[1,3]*sd_b555/PCA_R
  $vectors[3,3]/sd_PIM*mean(PIM)+-PCA_R
  $vectors[2,3]*sd_b555/PCA_R$vectors[3,3]/sd_
  POM*mean(POM)+mean(b555)
slope_R.PIM=-PCA_R$vectors[1,3]*sd_b555/
  PCA_R$vectors[3,3]/sd_PIM
slope_R.POM=-PCA_R$vectors[2,3]*sd_b555/
  PCA_R$vectors[3,3]/sd_POM
intercept_R
slope_R.PIM
slope_R.POM

#Generate bootstrap percentile intervals, presented in section 4f, for slope estimates and display #results.

set.seed(1234)
boots=10000
slopePIM_S.boot <- numeric(boots)
slopePOM_S.boot <- numeric(boots)
slopePIM_R.boot <- numeric(boots)
slopePOM_R.boot <- numeric(boots)
for(i in 1:boots)
{
index = 1:n
bootindex = sample(index, n, replace=T)
bootsample = combined[bootindex,]
Sb=cov(bootsample)
Rb=cor(bootsample)
PCA_Sb=eigen(Sb)
PCA_Rb=eigen(Rb)
slopePIM_S.boot[i]=-PCA_Sb$vectors[1,3]/PCA_Sb
  $vectors[3,3]
slopePOM_S.boot[i]=-PCA_Sb$vectors[2,3]/
  PCA_Sb$vectors[3,3]
slopePIM_R.boot[i]=-PCA_Rb$vectors[1,3]*sd
  (b555)/PCA_Rb$vectors[3,3]/sd(PIM)
slopePOM_R.boot[i]=-PCA_Rb$vectors[2,3]*sd
  (b555)/PCA_Rb$vectors[3,3]/sd(POM)
}

#Compute lower and upper confidence limits and display results

L95_S.PIM=quantile(slopePIM_S.boot,0.025)
U95_S.PIM=quantile(slopePIM_S.boot,0.975)
L95_S.POM=quantile(slopePOM_S.boot,0.025)
U95_S.POM=quantile(slopePOM_S.boot,0.975)
slope_S.PIM
L95_S.PIM
U95_S.PIM
slope_S.POM
L95_S.POM
U95_S.POM
L95_R.PIM=quantile(slopePIM_R.boot,0.025)
U95_R.PIM=quantile(slopePIM_R.boot,0.975)
L95_R.POM=quantile(slopePOM_R.boot,0.025)
U95_R.POM=quantile(slopePOM_R.boot,0.975)
```

slope_R.PIM
L95_R.PIM
U95_R.PIM
slope_R.POM
L95_R.POM
U95_R.POM

REFERENCES

Anderson, T. W., 1963: Asymptotic theory for principal component analysis. *Ann. Math. Stat.,* **34,** 122–148, doi:10.1214/aoms/1177704248.

——, 1984: Estimating linear statistical relationships. *Ann. Stat.,* **12,** 1–45.

Babin, M., A. Morel, V. Fournier-Sicre, F. Fell, and D. Stramski, 2003: Light scattering properties of marine particles in coastal and open ocean waters as related to the particle mass concentration. *Limnol. Oceanogr.,* **48,** 843–859, doi:10.4319/lo.2003.48.2.0843.

Banse, K., 1977: Determining the carbon-to-chlorophyll ratio of natural phytoplankton. *Mar. Biol.,* **41,** 199–212, doi:10.1007/BF00394907.

Clarke, A. J., and S. Van Goder, 2013: On fitting a straight line to data when the ''noise'' in both variables is unknown. *J. Atmos. Oceanic Technol.,* **30,** 151–158, doi:10.1175/JTECH-D-12-00067.1.

Fichot, C. G., S. Sathyendranath, and W. L. Miller, 2008: Sea*UV* and SeaUV_C: Algorithms for the retrieval of UV/visible diffuse attenuation coefficients from ocean color. *Remote Sens. Environ.,* **112,** 1584–1602, doi:10.1016/j.rse.2007.08.009.

Gallie, E. A., and P. A. Murtha, 1992: Specific absorption and backscattering spectra for suspended mineral and chlorophyll a in Chilko Lake, British Columbia. *Remote Sens. Environ.,* **39,** 103–118, doi:10.1016/0034-4257(92)90130-C.

Ikeda, T., 1970: Relationships between respiration rate and body size in marine plankton animals as a function of the temperature of habitat. *Bull. Fac. Fish. Hokkaido Univ.,* **21,** 91–112.

Jolicoeur, P., 1968: Interval estimation of the slope of the major axis of a bivariate normal distribution in the case of a small sample. *Biometrics,* **24,** 679–682, doi:10.2307/2528326.

——, 1975: Linear regression in fishery research: Some comments. *J. Fish. Res. Board Can.,* **32,** 1491–1494, doi:10.1139/f75-171.

Kendall, M. G., and A. Stuart, 1977: *Inference and Relationship.* Vol. 2, *The Advanced Theory of Statistics,* 4th ed. Griffin, 224 pp.

Kermack, K. A., and J. B. S. Haldane, 1950: Organic correlation and allometry. *Biometrika,* **37,** 30–41, doi:10.1093/biomet/37.1-2.30.

Laws, E. A., 1997: *Mathematical Models for Oceanographers.* John Wiley & Sons, 343 pp.

——, and J. W. Archie, 1981: Appropriate use of regression analysis in marine biology. *Mar. Biol.,* **65,** 13–16, doi:10.1007/BF00397062.

Pearson, K., 1901: On lines and planes of closest fit to systems of points in space. *Philos. Mag.,* **2,** 559–572, doi:10.1080/14786440109462720.

R Core Team, 2012: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Available online at http://www.R-project.org/.]

Rencher, A. C., 2002: *Methods of Multivariate Analysis.* 4th ed. John Wiley & Sons, 708 pp.

Ricker, W. E., 1973: Linear regressions in fishery research. *J. Fish. Res. Board Can.,* **30,** 409–434, doi:10.1139/f73-072.

Schlesinger, W. H., J. J. Cole, A. C. Finzi, and E. A. Holland, 2011: Introduction to coupled biogeochemical cycles. *Front. Ecol. Environ.,* **9,** 5–8, doi:10.1890/090235.

Sokal, R. R., and F. J. Rohlf, 1995: *Biometry: The Principles and Practice of Statistics in Biological Research.* 3rd ed. W.H. Freeman & Company, 885 pp.

Sprent, P., and G. R. Dolby, 1980: The geometric mean functional relationship. *Biometrics,* **36,** 547–550, doi:10.2307/2530224.

Stavn, R. H., 2012: Mass-specific scattering cross sections of suspended sediments and aggregates: Theoretical limits and applications. *Opt. Express,* **20,** 201–219, doi:10.1364/OE.20.000201.

——, and S. J. Richter, 2008: Biogeo-optics: Particle optical properties and the partitioning of the spectral scattering coefficient of ocean waters. *Appl. Opt.,* **47,** 2660–2679, doi:10.1364/AO.47.002660.

Steele, J. H., and I. E. Baird, 1961: Relations between primary production, chlorophyll and particulate carbon. *Limnol. Oceanogr.,* **6,** 68–78, doi:10.4319/lo.1961.6.1.0068.

Sverdrup, H., 1916: Druckgradient, Wind und Reibung an der Erdoberfläche. *Ann. Hydrogr. Marit. Meteor.,* **44,** 413–427.

Tett, P., J. C. Cottrell, D. O. Trew, and B. J. B. Wood, 1975: Phosphorus quota and the chlorophyll : carbon ratio in marine plankton. *Limnol. Oceanogr.,* **20,** 587–603, doi:10.4319/lo.1975.20.4.0587.

Warton, D. I., I. J. Wright, D. S. Falster, and M. Westoby, 2006: Bivariate line-fitting methods for allometry. *Biol. Rev.,* **81,** 259–291, doi:10.1017/S1464793106007007.