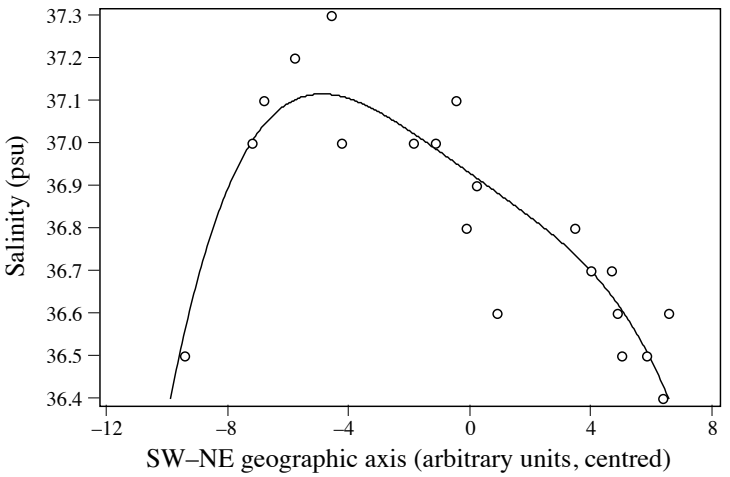

Regression equation:

	Regr. coeff.	Standard coeff.
b_0	36.929	0
$b_1(x)$	–0.0509	–0.9699
$b_2(x^4)$	–0.0001	–0.7261

Figure 10.9 Polynomial regression line describing the structure of salinity (psu: practical salinity units) in the Thau lagoon (Mediterranean Sea) along its main geographic axis on 25 October 1988.

is the projection of the positions of the sampling sites on the long axis of the lagoon, as determined by principal component analysis of the site coordinates. Being a principal component, variable x is centred. The other terms of an ordinary 6th-degree polynomial were computed from it. After stepwise selection, the model with the lowest AIC_c contained variables x, x^4 and x^5 ($AIC_c = –80.845$, $R^2_a = 0.815$); the regression parameters for x^4 and x^5 were not significant at the 0.05 level. Then, all possible models involving x, x^2, x^3, x^4 and x^5 were computed. The model with the largest number of significant regression coefficients contained variables x and x^4 ($AIC_c = –80.374$, $R^2_a = 0.792$). These results indicate that the model with three monomials (x, x^4 and x^5) is slightly better in terms of AIC_c and is thus the best-fitting model for the data. The line fitted to the second model, which is more parsimonious with only two explanatory variables (x and x^4), is shown in Fig. 10.9).

5 — Partial linear regression and variation partitioning

There are situations where two or more complementary sets of hypotheses may be invoked to explain the variation of an ecological variable. For example, the abundance of a species could vary as a function of biotic and abiotic factors. Regression modelling may be used to study one set of factors, or the other, or the two sets together. In most if not all cases involving field data (by opposition to experimental designs), there are correlations among variables across the two (or more) explanatory data sets. Partial regression is a way of estimating how much of the variation of the response variable can be attributed exclusively to one set once the effect of the other has been taken into account and controlled for. The purpose may be to estimate the amount of variation that can be attributed exclusively to one or the other set of explanatory variables and the amount explained jointly by the two explanatory data sets, or else to

estimate the vector of fitted values corresponding to the exclusive effect of one set of variables. When the objective is simply to assess the unique contribution of each explanatory variable, there is no need for partial regression analysis: the coefficients of multiple regression of the standardized variables already provide that information since they are *standard partial regression coefficients*.

Matrix of covariables

Consider three data sets. Vector **y** is the response variable whereas matrices **X** and **W** contain the explanatory variables. Assume that one wishes to model the relationship between **y** and **X**, while controlling for the effects of the variables in matrix **W**, which is called the *matrix of covariables*. The roles of **X** and **W** could of course be inverted.

Variation partitioning

Variation partitioning consists in apportioning the variation* of variable **y** among two or more explanatory data sets. This approach was first proposed by Mood (1969, 1971) and further developed by Borcard *et al*. (1992) and Peres-Neto *et al*. (2006). The method is described here for two explanatory data sets, **X** and **W**, but it can be extended to more explanatory matrices. When **X** and **W** contain random variables (defined at the beginning of Section 10.3), adjusted coefficients of determination (R^2_a, eq. 10.21) are used to compute the fractions following the method described below. Ordinary R^2 (eq. 10.20) are used instead of R^2_a when **X** and **W** represent ANOVA fixed factors coded into binary dummy variables or Helmert contrasts (Subsection 1.5.7).

Figure 10.10 sets a nomenclature, [a] to [d], for the fractions of variation that can be identified in **y**. Kerlinger & Pedhazur (1973) called this form of analysis "commonality analysis" by reference to the common fraction of variation (fraction [b] in Fig. 10.10) that two sets of explanatory variables may explain jointly. Partial regression assumes that the effects are linear and additive. There are two ways of carrying out the partitioning computations, depending on whether one wishes to obtain vectors of fitted values corresponding to fractions of variation, or simply estimate the amounts of variation corresponding to the fractions. In the description that follows, the fractions of variation are computed from R^2_a statistics.

(1) If one is interested in obtaining a partial regression equation and computing a vector of partial fitted values, one first computes the residuals of **y** on **W** (noted $\mathbf{y}_{\text{resl}\mathbf{W}}$) and the residuals of **X** on **W** (noted $\mathbf{X}_{\text{resl}\mathbf{W}}$):

Residuals of **y** on **W**: $$\mathbf{y}_{\text{resl}\mathbf{W}} = \mathbf{y} - \mathbf{W}\,[\mathbf{W}'\mathbf{W}]^{-1}\mathbf{W}'\,\mathbf{y}$$

Residuals of **X** on **W**: $$\mathbf{X}_{\text{resl}\mathbf{W}} = \mathbf{X} - \mathbf{W}\,[\mathbf{W}'\mathbf{W}]^{-1}\mathbf{W}'\,\mathbf{X}$$

In both cases, the regression coefficients are computed here through eq. 2.19 in which **X** is replaced by **W**. QR decomposition (see Section 10.7), which is also used in some

* The term *variation*, a less technical and looser term than *variance*, is used because one is partitioning the total sum of squared deviations of **y** from its mean (total SS). In variation partitioning, there is no need to divide the total SS of **y** by its degrees of freedom to obtain the variance s^2_y (eq. 4.3).

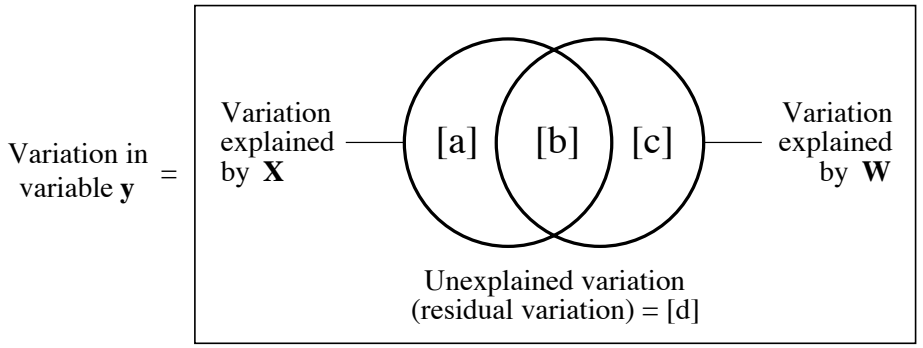

Figure 10.10 Partition of the variation of a response variable **y** among two sets of explanatory variables, **X** and **W**. The rectangle represents 100% of the variation in **y**. Fraction [b] is the intersection (*not* the interaction) of the variation explained by linear models of **X** and **W**. Adapted from Legendre (1993).

situations in Subsection 11.1, e.g. in Table 11.5, offers another way of computing regression equations.

Then, two computation methods are available: one can either

(1.1) regress $\mathbf{y}_{\text{resl}\mathbf{W}}$ on $\mathbf{X}_{\text{resl}\mathbf{W}}$,

(1.2) or regress **y** on $\mathbf{X}_{\text{resl}\mathbf{W}}$. The same partial regression coefficients are obtained in both cases, as will be verified in the numerical example below. Between calculation methods, the vectors of fitted values only differ by the values of the intercepts. The R^2 of analysis 1.1 is the partial R^2 whereas that of analysis 1.2 is the semipartial R^2; their square roots are the partial and semipartial correlation coefficients (Box 4.1).

(2) If one is interested in estimating the fractions resulting from partitioning the variation of vector **y** among the explanatory data sets **X** and **W**, there is a simple way to obtain the information, considering the ease with which multiple regressions can be computed using R or commercial statistical packages:

• Compute the multiple regression of **y** against **X** and **W** together. The corresponding R^2_a measures the fraction of information [a + b + c], which is the sum of the fractions of variation [a], [b], and [c] defined in Fig. 10.10. For the example data set (below), $R^2 = 0.5835$, so $R^2_a = 0.3913 = [a + b + c]$. The vector of fitted values corresponding to fraction [a + b + c], which is required to plot Fig. 10.13 (below), is also computed.

• Compute the multiple regression of **y** against **X**. The corresponding R^2_a measures [a + b], which is the sum of the fractions of variation [a] and [b]. For the example data,

$R^2 = 0.4793$, so $R^2_a = 0.3817 = [a + b]$. The vector of fitted values corresponding to fraction [a + b], which is required to plot Fig. 10.13, is also computed.

• Compute the multiple regression of **y** against **W**. The corresponding R^2_a measures [b + c], which is the sum of the fractions of variation [b] and [c]. For the example data, $R^2 = 0.3878$, so $R^2_a = 0.2731 = [b + c]$. The vector of fitted values corresponding to fraction [b + c], which is required to plot Fig. 10.13, is also computed.

• If needed, fraction [d] may be computed by subtraction. For the example, it is equal to 1 – [a + b + c], or 1 – 0.3913 = 0.6087.

As explained in Subsection 10.3.3, the adjusted R-square, R^2_a (eq. 10.21), is an unbiased estimator of the real contribution of a set of random variables **X** to the explanation of **y**. Following Peres-Neto *et al*. (2006), the values of the individual fractions [a], [b], and [c] must be computed by combining the R^2_a values obtained from the three multiple regressions that produced fractions [a + b + c], [a + b], and [b + c]:

• fraction [a] is computed by subtraction, using the R^2_a values: [a] = [a + b + c] – [b + c];

• likewise, fraction [c] is computed by subtraction, using the R^2_a values: [c] = [a + b + c] – [a + b];

• fraction [b] is also obtained by subtraction, using the R^2_a values, in the same way as the quantity B used for comparing two qualitative descriptors in Section 6.2:

$$[b] = [a + b] + [b + c] - [a + b + c] \quad or \quad [b] = [a + b] - [a] \quad or \quad [b] = [b + c] - [c]$$

Negative [b]

Fraction [b] may be negative. As such, it is not a rightful measure of variance; this is another reason why it is referred to by the looser term *variation*. A negative fraction [b] indicates that two variables (or groups of variables **X** and **W**), together, explain **y** better than the sum of the individual effects of these variables. This can happen: see Numerical examples 2 and 3. Fraction [b] is the *intersection* of the variation explained by linear models of **X** and **W**. It is *not an interaction* in the ANOVA sense.

Vectors of fitted values corresponding to fractions [a] and [c] can be computed using partial regression, as explained above, while the vector of residuals of the regression equation that uses all predictors corresponds to fraction [d]. No fitted vector can be estimated for fraction [b], however, because no partial regression model can be written for that fraction. No degrees of freedom are attached to fraction [b]; hence [b] cannot be tested for significance.

Selection of explanatory variables

If a selection procedure (backward, forward, stepwise; Subsection 10.3.3) is used, it must be applied to data matrices **X** and **W** separately, before partitioning, in order to preserve fraction [b] of the partition. Applying the selection to matrices **X** and **W** combined could result in the elimination of variables from one or both matrices because they are correlated with variables in the other matrix, thereby reducing or eliminating fraction [b].

Table 10.6 Data collected at 20 sites in the Thau coastal lagoon on 25 October 1988. There are two bacterial response variables (Bna and Ma), three environmental variables (NH_4, phaeopigments, and bacterial production), and three spatial variables (the X and Y geographic coordinates measured with respect to arbitrary axes and centred on their respective means, plus the quadratic monomial X^2). The variables are further described in the text. The code names of these variables in the present section are y, x_1 to x_3, and w_1 to w_3, respectively.

Site No.	Bna	Ma y	NH_4 x_1	Phaeo. *a* x_2	Prod. x_3	X w_1	Y w_2	X^2 w_3
1	4.615	10.003	0.307	0.184	0.274	–8.75	3.7	76.5625
2	5.226	9.999	0.207	0.212	0.213	–6.75	2.7	45.5625
3	5.081	9.636	0.140	0.229	0.134	–5.75	1.7	33.0625
4	5.278	8.331	1.371	0.287	0.177	–5.75	3.7	33.0625
5	5.756	8.929	1.447	0.242	0.091	–3.75	2.7	14.0625
6	5.328	8.839	0.668	0.531	0.272	–2.75	3.7	7.5625
7	4.263	7.784	0.300	0.948	0.460	–1.75	0.7	3.0625
8	5.442	8.023	0.329	1.389	0.253	–0.75	–0.3	0.5625
9	5.328	8.294	0.207	0.765	0.235	0.25	–1.3	0.0625
10	4.663	7.883	0.223	0.737	0.362	0.25	0.7	0.0625
11	6.775	9.741	0.788	0.454	0.824	0.25	2.7	0.0625
12	5.442	8.657	1.112	0.395	0.419	1.25	1.7	1.5625
13	5.421	8.117	1.273	0.247	0.398	3.25	–4.3	10.5625
14	5.602	8.117	0.956	0.449	0.172	3.25	–2.3	10.5625
15	5.442	8.487	0.708	0.457	0.141	3.25	–1.3	10.5625
16	5.303	7.955	0.637	0.386	0.360	4.25	–5.3	18.0625
17	5.602	10.545	0.519	0.481	0.261	4.25	–4.3	18.0625
18	5.505	9.687	0.247	0.468	0.450	4.25	–2.3	18.0625
19	6.019	8.700	1.664	0.321	0.287	5.25	–0.3	27.5625
20	5.464	10.240	0.182	0.380	0.510	6.25	–2.3	39.0625

Numerical example 1. The example data set (Table 10.6) is from the ECOTHAU research program mentioned in the numerical example of Subsection 10.3.4 (Amanieu *et al*., 1989). It contains two bacterial variables (Bna, the concentration of colony-forming units of aerobic heterotrophs growing on bioMérieux nutrient agar, with low NaCl concentration; and Ma, the concentration of aerobic heterotrophs growing on marine agar with a salt content of 34 gL^{-1}); three environmental variables (NH_4 in the water column, in $\mu molL^{-1}$; phaeopigments from degraded chlorophyll *a*, in μgL^{-1}; and bacterial production, determined by incorporation of tritiated thymidine in bacterial DNA, in $nmolL^{-1}d^{-1}$); and three spatial variables of the sampling sites on the nodes of an arbitrarily located grid (the X and Y geographic coordinates, in km, each centred on its mean, and the quadratic monomial X^2, which was found to be important for explaining the response variables). All bacterial and environmental variables were log-transformed using $\log_e(x + 1)$. One of the bacterial variables, Ma, is used here as the response variable **y**; the three environmental variables form the matrix of explanatory variables **X**; the three spatial variables make up matrix **W** of the covariables. Table 10.6 will be used again in

Section 13.4. A multiple regression of **y** against **X** and **W** together was computed first as a reference. The regression equation was the following:

$$\hat{y} = 9.64 - 0.90x_1 - 1.34x_2 + 0.54x_3 + 0.10w_1 + 0.14w_2 + 0.02w_3$$
$$(R^2 = 0.5835;\ R^2_a = 0.3913 = [\text{a} + \text{b} + \text{c}])$$

The adjusted coefficient of determination (R^2_a) is an unbiased estimate of the proportion of the variation of **y** explained by the regression model containing the 6 explanatory variables; it corresponds to fraction [a+b+c] in the partitioning table below and to the sum of fractions [a], [b] and [c] in Fig. 10.10. The vector of fitted values was also computed; after centring, this vector will be plotted as fraction [a + b + c] in Fig. 10.13. Since the total sum of squares in **y** is 14.9276 [$SS = s^2_y \times (n-1)$], the R^2 allowed the computation of the sum of squares corresponding to the vector of fitted values: $SS(\hat{\mathbf{y}}) = 14.9276 \times 0.5835 = 8.7109$. This value can also be obtained by computing directly the sum of squared deviations about the mean of the values in the fitted vector $\hat{\mathbf{y}}$.

For calculation of the partial regression equation using method 1.1, the residuals* of the regression of **y** on **W** were computed. One way is to use the following equation, which requires adding a column of "1" to matrix **W** in order to estimate the regression intercept:

$$\mathbf{y}_{\text{resl}\mathbf{W}} = \mathbf{y} - \mathbf{W}\,[\mathbf{W}'\mathbf{W}]^{-1}\mathbf{W}'\,\mathbf{y}$$

The residuals of the regressions of **X** on **W** were computed in the same way:

$$\mathbf{X}_{\text{resl}\mathbf{W}} = \mathbf{X} - \mathbf{W}\,[\mathbf{W}'\mathbf{W}]^{-1}\mathbf{W}'\,\mathbf{X}$$

Then, vector $\mathbf{y}_{\text{resl}\mathbf{W}}$ was regressed on matrix $\mathbf{X}_{\text{resl}\mathbf{W}}$ with the following result:

regression equation: $\hat{y} = 0 - 0.90x_{r(\mathbf{W})1} - 1.34x_{r(\mathbf{W})2} + 0.54x_{r(\mathbf{W})3}$ $(R^2 = 0.3197)$

The value $R^2 = 0.3197$ is the partial R^2. In its calculation, the denominator is the sum of squares corresponding to fractions [a] and [d], as shown for the partial correlation coefficient in Box 4.1.

For calculation through method 1.2, **y** was regressed on matrix $\mathbf{X}_{\text{resl}\mathbf{W}}$ with the following result:

regression equation: $\hat{y} = 8.90 - 0.90x_{r(\mathbf{W})1} - 1.34x_{r(\mathbf{W})2} + 0.54x_{r(\mathbf{W})3}$ $(R^2 = 0.1957)$

The value $R^2 = 0.1957$ is the semipartial R^2. The semipartial R^2 is the square of the semipartial correlation defined in Box 4.1. It represents the fraction of the total variation of **y** explained by the partial regression equation because, in its calculation, the denominator is the total sum of squares of the response variable **y**, [a+b+c+d]. That value is shown in the variation partitioning table below, but it will not be used to compute the individual fractions of variation.

Note that the three regression coefficients for the three *x* variables in the last equation are exactly the same as in the two previous equations; only the intercepts differ. This gives substance to the statement of Subsection 10.3.3 that regression coefficients obtained in multiple

* In the R language, regression residuals can be computed using ***residuals(lm())***.

linear regression are *partial regression coefficients* in the sense of the present subsection. Between calculation methods, the vectors of fitted values only differ by the value of the intercept of the regression of **y** on $\mathbf{X}_{\text{resl}\mathbf{W}}$, 8.90, which is also the mean of **y**. The centred vector of fitted values will be plotted as fraction [a] in Fig. 10.13.

The calculation of partial regression can be done in the opposite way, regressing **y** on **W** while controlling for the effects of **X**. First, $\mathbf{y}_{\text{resl}\mathbf{X}}$ and $\mathbf{W}_{\text{resl}\mathbf{X}}$ were computed. Then, for method 1.1, $\mathbf{y}_{\text{resl}\mathbf{X}}$ was regressed on $\mathbf{W}_{\text{resl}\mathbf{X}}$ with the following result:

regression equation: $\hat{y} = 0 - 0.10\,w_{r(\mathbf{X})1} - 0.14\,w_{r(\mathbf{X})2} + 0.02\,w_{r(\mathbf{X})3}$ $(R^2 = 0.2002)$

where $R^2 = 0.2002$ is the partial R^2. For method 1.2, **y** was regressed on $\mathbf{W}_{\text{resl}\mathbf{X}}$ with the following result:

regression equation: $\hat{y} = 8.90 - 0.10\,w_{r(\mathbf{X})1} - 0.14\,w_{r(\mathbf{X})2} + 0.02\,w_{r(\mathbf{X})3}$ $(R^2 = 0.1043)$

where $R^2 = 0.1043$ is the semipartial R^2, shown in the variation partitioning table below, but not used to compute the individual fractions of variation.

Again, the three regression coefficients in these partial regression equations are exactly the same as in the first regression equation of this example; only the intercepts differ. Between calculation methods, the vectors of fitted values only differ by the value of the intercept of the regression of **y** on $\mathbf{X}_{\text{resl}\mathbf{W}}$, 8.90, which is also the mean of **y**. The centred vector of fitted values will be plotted as fraction [c] in Fig. 10.13.

To estimate fraction [a + b] of Fig. 10.10, the multiple regression of **y** on the three original (non-residualized) variables in **X** was computed. The regression equation was:

$$\hat{y} = 10.20 - 0.93x_1 - 2.02x_2 + 0.89x_3 \qquad (R^2 = 0.4793;\ R^2_a = 0.3817 = [\text{a} + \text{b}])$$

The value $R^2_a = 0.3817$ is an unbiased estimate of the fraction of the variation of **y** accounted for by the linear model of the three explanatory variables **X**. The vector of fitted values was computed; after centring, this vector will be plotted as fraction [a + b] in Fig. 10.13.

To obtain fraction [b + c] of Fig. 10.10, the multiple regression of **y** on the three original (non-residualized) variables in **W** was computed. The regression equation was:

$$\hat{y} = 8.32 + 0.09w_1 + 0.10w_2 + 0.03w_3 \qquad (R^2 = 0.3878;\ R^2_a = 0.2731 = [\text{b} + \text{c}])$$

The value $R^2_a = 0.2731$ is the unbiased estimation of the fraction of the variation of **y** accounted for by the linear model of the three explanatory variables **W**. The vector of fitted values was computed; after centring, this vector will be plotted as fraction [b + c] in Fig. 10.13.

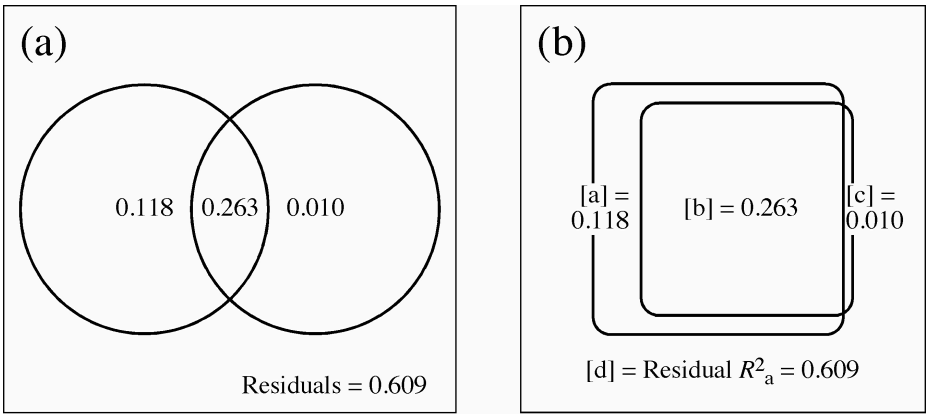

Figure 10.11 Venn diagram illustrating the results of variation partitioning of the numerical example. (a) Diagram drawn by the plotting function ***plot.varpart()*** of the VEGAN package. The circles are of equal sizes despite differences in the corresponding R^2_a. (b) Prior to publication of the partitioning results, the diagram can be redrawn, here using rounded rectangles, to better represent the relative fraction sizes with respect to the size of the outer rectangle, which represents the total variation in the response data. The fractions are identified by letters [a] to [d]; the value next to each identifier is the adjusted R^2 (R^2_a). Rectangle sizes are approximate.

Following the fraction nomenclature convention set in Fig. 10.10, the variation partitioning results were assembled in the following table (rounded values):

Fractions of variation	Sums of squares (SS)	Proportions of variation of y (R^2)	Adjusted R^2 (R^2_a)
[a + b]	7.1547	0.4793	0.3817
[b + c]	5.7895	0.3878	0.2731
[a + b + c]	8.7109	0.5835	0.3913
[a]	2.9213	*0.1957*	0.1183
[b]	4.2333	*0.2836*	0.2634
[c]	1.5562	*0.1043*	0.0097
Residuals = [d]	6.2167	*0.4165*	0.6087
[a + b + c + d]	14.9276	1.0000	1.0000

The partitioning results are illustrated as Venn diagrams in Fig. 10.11*. In Chapter 11, Fig. 11.6 shows partitioning results for multivariate response data involving three explanatory matrices.

* A Venn diagram with proportional circle and intersection sizes can be obtained using function ***venneuler()*** of the same-name package (Section 10.7).

As mentioned at the beginning of this subsection, and following Peres-Neto *et al*. (2006), when **X** and **W** contain random variables, R^2_a values corresponding to [a + b + c], [a + b], and [b + c] are used to compute, by subtraction, the fractions [a] to [d] shown in column 4 of the table. R^2_a provides unbiased estimates of the contributions of the explanatory data sets **X** and **W** to **y** when **X** and **W** contain random variables. The adjusted fractions [a], [b], and [c] cannot be directly computed using the non-adjusted fractions computed from non-adjusted R^2 coefficients, shown in italics in the 3*rd* column. When *n* is small as in this example, the estimated fractions computed from R^2_a may be very different from the fractions computed from R^2 values.

Ordinary R^2 (3*rd* column) are used to compute the fractions (values in italics) when **X** and **W** represent ANOVA fixed factors coded into dummy variables. When these values are required, they can be calculated by subtraction from the R^2 values in the first three rows of the table: R^2[a] = R^2[a+b+c] – R^2[b+c] = 0.1957 (which is equal to the R^2 of the partial regression equation computed above through method 1.2); R^2[c] = R^2[a+b+c] – R^2[a+b] = 0.1043 (which is equal to the R^2 of the partial regression equation computed above through method 1.2); R^2[b] = R^2[a+b] + R^2[b+c] – R^2[a+b+c] = 0.2836 (this value can only be obtained by subtraction). The sums of squares in the 2nd column of the table are obtained by multiplying these R^2 values by the total sum of squares in **y**, which is 14.9276.

The *partial correlation coefficient* between **y** and matrix **X** while controlling for the effect of **W** can be obtained from the values [a] and [d] in the column "Sums of squares" of the table, as explained in Box 4.1 of Section 4.5:

$$r_{\mathbf{yX.W}} = \sqrt{\frac{[a]}{[a+d]}} = \sqrt{\frac{2.9213}{2.9213 + 6.2167}} = 0.5654$$

This value is not the same as the *semipartial* R^2, which is computed as follows (Box 4.1):

$$r_{\mathbf{y(X.W)}} = \sqrt{\frac{[a]}{[a+b+c+d]}} = \sqrt{\frac{2.9213}{14.9276}} = 0.4424$$

Tests of significance of the fractions

If the conditions of homoscedasticity and normality of the residuals are satisfied, the fractions (with the exception of [b]) can be tested for significance through parametric tests. For fractions [a + b + c], [a + b], and [b + c], one can use the results of the parametric tests produced by the statistical software. For fractions [a] and [c], one must construct a *F*-statistic as in eq. 11.22, using the sum of squares corresponding to fraction [a] (symbol: SS[a]) or [c] (symbol: SS[c]) in the numerator, and the residual sum of squares corresponding to [d] (symbol: SS[d]) in the denominator, together with appropriate numbers of degrees of freedom. The test statistic for fraction [a], for example, is constructed as follows:

$$F_{[a]} = \frac{SS[a]/m}{SS[d]/(n-m-q-1)}$$

where *m* is the number of explanatory variables in set **X** and *q* is the number of covariables in set **W**. In the parametric framework, the statistic is tested against the

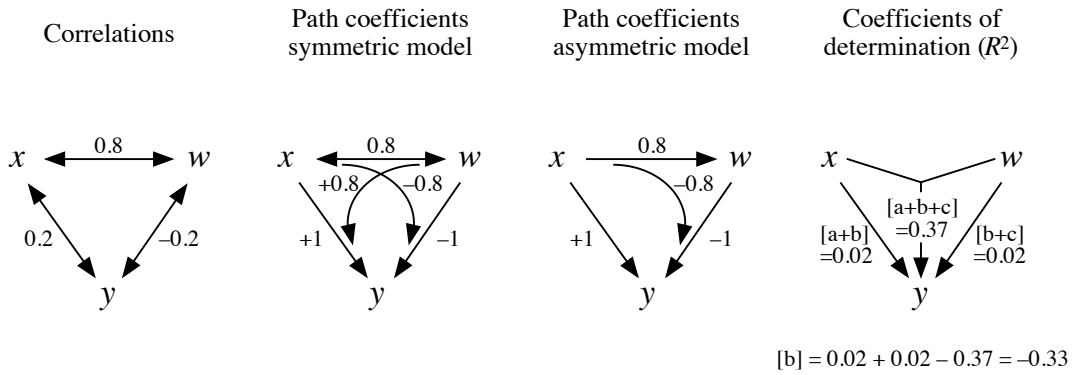

Figure 10.12 Correlations, path coefficients, and coefficients of determination for Numerical example 2.

F–distribution with m and $(n - m - q - 1)$ degrees of freedom. An example of that *F*-statistic for the test of partial or semipartial correlation coefficients is given in Box 4.1 for the simple case where there is a single variable in **X** and **W**.

If the conditions of homoscedasticity or normality of the residuals are not satisfied, one can use permutation tests to obtain p-values. Permutation of the raw data is used to test fractions [a + b + c], [a + b], and [b + c]. To test fractions [a] and [c], permutation of the residuals of a null or full model should be used (Anderson & Legendre, 1999). These permutation methods are described in Subsection 11.1.8.

Numerical example 2. This example illustrates the appearance of a negative fraction [b] when there are strong direct effects of opposite signs of x and w on y and a strong correlation between x and w (non-orthogonality). For three variables measured over 50 objects, the following correlations are obtained: $r(x, w) = 0.8$, $r(y, x) = 0.2$ and $r(y, w) = -0.2$; y, x, and w have the same meaning as in the previous numerical example. $r(y, x)$ and $r(y, w)$ are not statistically significant at the $\alpha = 0.05$ level. Referring to Section 10.4, one may use path analysis to compute the direct and indirect causal covariation relating the explanatory variables x and w to the response variable y. One can also compute the coefficient of determination of the model $y = f(x, w)$; its value is $R^2 = 0.40$. From these values, the partition of the variation of y can be achieved: R^2 of the whole model = 0.40, R^2_a = [a + b + c] = 0.37447; $r^2(w, y) = 0.04$, R^2_a = [a + b] = 0.02; $r^2(x, y) = 0.04$, R^2_a = [b + c] = 0.02. Hence, [b] = [a + b] + [b + c] – [a + b + c] = –0.33447, [a] = [a + b] – [b] = 0.35447, and [c] = [b + c] – [b] = 0.35447. How is that possible?

Carrying out path analysis (Fig. 10.12), and assuming a symmetric model of relationships (i.e. w affects x and x affects w), the direct effect of x on y, $p_{xy} = 1.0$, is positive and highly significant, but it is counterbalanced by a strong negative indirect covariation of –0.8 going through w. In the same way, $p_{wy} = -1.0$ (which is highly significant), but this direct effect is counterbalanced by a strong positive indirect covariation of +0.8 going through x. As a result, and although they both have the maximum possible value of 1.0 for direct effects on the

response variable y, both w and x turn out to have non-significant total correlations with y. In the present variation partitioning model, this translates into small adjusted amounts of explained variation $[a + b] = 0.02$ and $[b + c] = 0.02$, and a negative value for fraction [b]. If an asymmetric model of relationship had been assumed (e.g. w affects x but x does not affect w), essentially the same conclusion would have been reached from path analysis.

Numerical example 3. Another situation can give rise to a negative fraction [b], i.e. when there is no linear correlation between y and one of the explanatory variables, e.g. $r(y, x) = 0.0$, but the other two correlations differ from 0, e.g. $r(y, w) = 0.5$ and $r(x, w) = 0.5$. For this example, assuming again $n = 50$, we find $[a + b + c] = 0.30497$, $[a + b] = -0.02083$, and $[b + c] = 0.23438$ (computed from the R^2_a coefficients), so that $[b] = -0.09142$. The partial explanation of the variation of y provided by x, estimated by the partial regression or partial correlation coefficient, is not zero and may be significant in the statistical sense: using path analysis (Section 10.4) for this example, the direct effect of x on y is $p_{xy} = -0.33333$ ($p = 0.019$, which is significant) and the indirect effect is 0.33333, these two effects summing to zero. The direct effect of w on y is $p_{wy} = 0.66667$ and its indirect effect is -0.16667. The negative [b] fraction indicates that x and w, together, explain the variation of y better than the sum of the individual effects of these variables. The signs of the regression coefficients (path coefficients) actually vary depending on the signs of the correlations $r(y, w)$ and $r(x, w)$.

The above decomposition of the variation of a response vector **y** between two sets of explanatory variables **X** and **W** was described by Whittaker (1984) for the simple case where there is a single regressor in each set **X** and **W**. Whittaker showed that the various fractions of variation may be represented as vectors in space, and that the value of fraction [b] [noted G(12:) by Whittaker, 1984] is related to the angle θ between the two regressors through the following formula:

$$1 - 2\cos^2(\theta/2) \leq [b] \leq 2\cos^2(\theta/2) - 1 \qquad \textbf{(10.25)}$$

Fraction [b] for orthogonal regressors

θ is related to the coefficient of linear correlation (eq. 10.4). This formula has three interesting properties. (1) If the two regressors are orthogonal ($r = 0$), then $2\cos^2(\theta/2) = 1$, so that $0 \leq [b] \leq 0$ and consequently $[b] = 0$. Turning the argument around, the presence of a non-zero fraction [b] indicates that the two explanatory variables are not orthogonal. There are also instances where [b] is zero with two non-orthogonal regressors; a simple example is when the two regressors are uncorrelated with **y** and explain none of its variation. (2) If the two regressors are identical, or at least pointing in the same direction ($\theta = 0°$), then $-1 \leq [b] \leq 1$. It follows that the proportion of variation of **y** that is accounted for by either regressor (fraction [b]) may be, in some cases, as large as 1, i.e. 100%. (3) The formula allows for negative values of [b], as shown in Numerical example 2.

In conclusion, fraction [b] represents the fraction of variation of **y** that may indifferently be attributed to **X** or **W**. The interpretation of a negative [b] is that the two processes, represented in the analysis by data sets **X** and **W**, are competitive; in other words, they have opposite effects, one process hindering the contribution of the other in the joint regression model. One could use eq. 6.15, $S = [b]/[a + b + c]$, to quantify how similar **X** and **W** are in explaining **y**. Whittaker (1984) also suggested that if **X** and **W** represent two factors of an experimental design, [b] may be construed as a