included in explanatory matrix **X**. As in simple linear regression, where the coefficient of determination is r^2 (eq. 10.7), $R^2_{\mathbf{y}|\mathbf{X}}$ is the regression sum of squares (SS) divided by the total sum of squares (total SS, TSS), or the one-complement of the ratio of the sum of squared residuals (residual sum of squares, RSS) to the total sum of squares (TSS):

$$R^2_{\mathbf{y}|\mathbf{X}} = \frac{\text{regression SS}}{\text{total SS}} = \frac{\Sigma (\hat{y}_i - \bar{y})^2}{\Sigma (y_i - \bar{y})^2} = 1 - \frac{\text{RSS}}{\text{TSS}} \qquad \textbf{(10.20)}$$

The expected value of R^2 in a regression involving m random predictors is not 0 but $m/(n-1)$, as explained below. As a consequence, if **X** contains $m = (n-1)$ predictors that are linearly unrelated to the response variable y, for example m columns of random numbers, $R^2 = 1$ even though the explanatory variables explain none of the variation of y. For that reason, R^2 cannot be interpreted as a correct (i.e. unbiased) estimate of the proportion of variation of y explained by **X**.

Three useful statistics can, however, be derived from R^2. They serve distinct purposes in regression analysis.

Adjusted R^2

1. The *adjusted coefficient of multiple determination* R^2_a or *adjusted* R^2 (Ezekiel, 1930), provides an unbiased estimate of the proportion of variation of y explained by **X**. The formula takes into account the numbers of degrees of freedom (d.f.) of the numerator and denominator portions of R^2:

$$R^2_a = 1 - \frac{\text{residual mean square}}{\text{total mean square}} = 1 - (1 - R^2_{\mathbf{y}|\mathbf{X}}) \left(\frac{\text{total d.f.}}{\text{residual d.f.}} \right) \qquad \textbf{(10.21)}$$

- In ordinary multiple regression, the total degrees of freedom of the F-statistic are $(n-1)$ and the residual d.f. are $(n-m-1)$, where n is the number of observations and m is the number of explanatory variables in the model (eq. 4.40).

- In multiple regression through the origin, where the intercept is forced to zero, the total degrees of freedom of the F-statistic are n and the residual d.f. are $(n-m)$.

These same degrees of freedom are used in eq. 10.21. The logic of this adjustment is the following: in ordinary multiple regression, a random predictor explains on average a proportion $1/(n-1)$ of the response's variation, so that m random predictors explain together, on average, $m/(n-1)$ of the response's variation; in other words, the expected value of R^2 is $E(R^2) = m/(n-1)$. Applying eq. 10.21 to that value, where all predictors are random, gives $R^2_a = 0$. In regression through the origin, a random predictor explains on average a proportion $1/n$ of the response's variation, so that m random predictors explain together, on average, m/n of the response's variation, and $R^2 = m/n$. Applying eq. 10.21 to that case gives, again, $R^2_a = 0$.

R^2_a is a suitable measure of goodness of fit for comparing the success of regression equations fitted to different data sets, with different numbers of objects and explanatory variables. Using simulated data with normal error, Ohtani (2000) has shown that R^2_a is an unbiased estimator of the contribution of a set of random

predictors **X** to the explanation of *y*. This adjustment may be too conservative when $m > n/2$ (Borcard *et al*., 2011); this is a rule of thumb rather than a statistical principle.

With real matrices of random variables (defined at the beginning of Section 10.3), when the explanatory variables explain no more of the response's variation than the same number of variables containing random numbers, the value of R^2_a is near zero; it can be negative on occasion. Contrary to R^2, R^2_a does not necessarily increase with the addition of explanatory variables to the regression model if these explanatory variables are linearly unrelated to *y*. R^2_a is a better estimate of the population coefficient of determination ρ^2 than R^2 (Zar, 1999, Section 20.3) because it is unbiased.

Healy (1984) pointed out that Ezekiel's (1930) adjusted R^2 equation (R^2_a, eq. 10.21) makes sense and should be used when **X** contains observed values of random variables. That is not the case for ANOVA fixed factors, which can be used in a multiple regression equation when they are recoded into binary dummy variables or Helmert contrasts (Subsection 1.5.7).

In canonical analysis (Chapter 11), the canonical R^2 is called the *bimultivariate redundancy statistic* (Miller & Farr, 1971), *canonical coefficient of determination*, or *canonical* R^2. Using numerical simulations, Peres-Neto *et al*. (2006) have shown that, in redundancy analysis (RDA, Section 11.1), for normally distributed data or Hellinger-transformed species abundances, the adjusted canonical R^2 (R^2_a, eq. 11.5), obtained by applying eq. 10.21 to the canonical R^2 ($R^2_{\mathbf{Y}|\mathbf{X}}$, eq. 11.4), produces unbiased estimates of the contributions of the variables in **X** to the explanation of a response matrix **Y**, just as in multiple regression. With simulated data, they also showed the artificial increase of R^2 as the number of unrelated explanatory variables in explanatory matrix **X** increases.

AIC, AIC_c

2. The *Akaike Information Criterion* (*AIC*) is a measure of the goodness of fit of the data to an estimated statistical model (Akaike, 1974). When comparing linear regression models, *AIC* is computed as follows (RSS, TSS: see eq. 10.20):

$$AIC = n \log_e\left(\frac{\text{RSS}}{n}\right) + 2k \qquad \textbf{(10.22)}$$

where k is the number of parameters, including the intercept, in the regression equation. Independence of the observations is assumed in the calculation of *AIC*, as well as normality of the residuals and homogeneity of their variances. The following formula is also found in the literature: $AIC = n \log_e((1 - R^2)/n) + 2k$. A constant, $n\log_e(\text{TSS})$, must be added to this formula to obtain eq. 10.22. Since *AIC* is used to compare different models of the same response data, either formula will identify the same model as the one that minimizes *AIC*.

The corrected form of *AIC*, abbreviated AIC_c (Hurvich & Tsai, 1993), is *AIC* with a second-order correction for small sample size:

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1} \qquad \textbf{(10.23)}$$

Burnham & Anderson (2002) strongly recommend using AIC_c rather than AIC when n is small or k is large. Because AIC_c converges towards AIC when n is large, AIC_c should be used with all sample sizes.

The AIC_c statistic is not the basis for a test of significance. It plays a different role than the F-test (below): it is used to compare models. For a given data set, several competing models may be ranked by AIC_c. The model with the *smallest* value of AIC_c is the best-fitting one, i.e. the most likely for the data. For example, in selection of explanatory variables, the model for which AIC_c is minimum is retained.

F-statistic

3. The F-statistic (see eq. 4.40) serves as the basis for the test of significance of the coefficient of multiple determination, R^2. A parametric test can be used if the regression residuals are normal. Otherwise, a permutation test should be used.

F-statistic for nested models

There is another way of comparing models statistically, but it is limited to nested models of the same response data. A model is nested in another if it contains one or several variables less than the reference model. The method consists in calculating the R^2 of the two linear models and computing a F-statistic to test the difference in R^2 between them. The F-statistic is computed as follows for two nested models, the most inclusive containing m_2 variables and the model nested into it containing m_1 variables:

$$F = \frac{(R^2_{y.1\ldots m_2} - R^2_{y.1\ldots m_1}) / (m_2 - m_1)}{(1 - R^2_{y.1\ldots m_2}) / (n - m_2 - 1)}$$

The difference in R^2 is tested for significance parametrically with $\nu_1 = (m_2 - m_1)$ and $\nu_2 = (n - m_2 - 1)$ degrees of freedom, or by permutation. This method can be used in forward selection or backward elimination. It is implemented, for example, in functions ***ordiR2step()*** of VEGAN and ***forward.sel()*** of PACKFOR (Subsection 11.1.10, paragraph 7), which can be used in models involving a single response variable y.

As a final note, it is useful to remember that several types of explanatory variables can be used in multiple regression:

Dummy variable regression

• Binary descriptors can be used as explanatory variables in multiple regression, together with quantitative variables. This means that multistate qualitative variables can also be used, insofar as they are recoded into binary dummy variables, as described in Subsection 1.5.7*. This case is referred to as *dummy variable regression*.

• Geographic information may be used in multiple regression models in different ways. On the one hand, latitude (Y) and longitude (X) information form perfectly valid quantitative descriptors if they are recorded as axes of a Cartesian plane. Geographic data in the form of degrees-minutes-seconds should, however, be recoded to decimal

* In R, qualitative multistate descriptors used as explanatory variables are automatically recoded into dummy variables by function ***lm()*** if they are identified as *factors* in the data frame.

form before they are used as explanatory variables in regression. The X and Y coordinates may be used either alone, or in the form of a polynomial (X, Y, X^2, XY, Y^2, etc.). Regression using such explanatory variables is referred to as *trend surface analysis* in Chapter 13. Spatial eigenfunctions, described in Chapter 14, are more sophisticated descriptions of geographic relationships among study sites; they can also be used as explanatory variables in regression.

• If replicate observations are available for each site, the grouping of observations, which is also a kind of geographic information, may be used in multiple regression as a qualitative multistate descriptor, recoded into a set of dummy variables.

• Finally, any analysis of variance may be reformulated as a linear regression analysis; actually, linear regression and ANOVA both belonging to the General Linear Model. Consider one-way ANOVA for instance: the classification criterion can be written as a multistate qualitative variable and, as such, recoded as a set of dummy variables (Subsection 1.5.7) on which multiple regression may be performed. The analysis of variance table obtained by multiple regression is identical to that produced by ANOVA. This equivalence is discussed in more detail by ter Braak & Looman (1987) in an ecological framework. Draper & Smith (1981) and Searle (1987) discuss in some detail how to apply multiple regression to various analysis of variance configurations. ANOVA by regression can be extended to cross-factor (two-way or multiway) ANOVA. How to carry out these analyses is described in Subsection 11.1.10, point 4, for the more general analysis of multivariate response data **Y** (MANOVA).

4 — Polynomial regression

Several solutions have been proposed to the problem of fitting, to a response variable y, a nonlinear function of a single explanatory variable x. An elegant and easy solution is to use a polynomial of x, whose terms are treated as so many explanatory variables in a multiple regression procedure. In this approach, y is modelled as a polynomial function of x:

Polynomial model

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \ldots + b_k x^k \qquad \textbf{(10.24)}$$

Such an equation is linear in its parameters (if one considers the terms $x^2, \ldots, x^k$ as so many explanatory variables), although the modelled response of y to the explanatory variable x is nonlinear. The degree of the equation, which is its highest exponent, determines the shape of the curve: each degree above 1 (straight line) and 2 (concave up or down) adds an inflexion point to the curve. Increasing the degree of the equation always increases its adjustment to the data (R^2). If one uses as many parameters b (including the intercept b_0) as there are data points, one can fit the data perfectly ($R^2 = 1$). However, the cost of that perfect fit is that there are no degrees of freedom left to test the relationship and, therefore, the "model" cannot be extended to other situations. Hence, a perfectly fitted model is useless. In any case, a high-degree polynomial would be of little interest in view of the principle of parsimony (Ockham's razor) discussed in Subsection 10.3.3, which states that the best model is the simplest