Coefficient of determination

$$R^2 = \frac{\text{regression SS}}{\text{total SS}} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\text{residual SS}}{\text{total SS}} \qquad \textbf{(10.19)}$$

R^2 measures the proportion of the variation of y about its mean that is explained by the regression equation.

Adjusted coefficient of determination

Another measure, the *adjusted coefficient of multiple determination* R^2_a (Ezekiel 1930), takes into account the respective numbers of degrees of freedom of the numerator and denominator of R^2:

$$R^2_a = 1 - \frac{\text{residual mean square}}{\text{total mean square}} = 1 - (1 - R^2)\left(\frac{\text{total d.f.}}{\text{residual d.f.}}\right) \qquad \textbf{(10.20)}$$

R^2_a is a suitable measure of goodness of fit for comparing 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 explanatory variables **X** to the explanation of **y**.

- In ordinary multiple regression, the total degrees of freedom (d.f.) 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 (Box 4.1).

- 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 $(n - m)$.

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, $R^2 = m/(n - 1)$. Applying eq. 10.20 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. 10.20 to that case gives, again, $R^2_a = 0$. With real tables of predictors, when the explanatory variables explain no more of the response's variation than random predictors would, 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. R^2_a is a better estimate than R^2 of the population coefficient of determination ρ^2 (Zar 1999, Section 20.3).

Adjusted bimultivariate redundancy statistic

In canonical analysis (Chapter 11), the canonical R^2 is called the *bimultivariate redundancy statistic* (Miller & Farr 1971) or canonical coefficient of determination. In the program CANOCO, the "trace statistic" for RDA is a non-adjusted canonical R^2. Using numerical simulations, Peres-Neto *et al.* (2006) have shown that, for normally distributed data or Hellinger-transformed species abundances in redundancy analysis (RDA, Section 11.1), the *adjusted bimultivariate redundancy statistic*, obtained by applying eq. 10.20 to the canonical R^2, produces unbiased estimates of the real contributions of the variables in **X** to the explanation of a response matrix **Y**.