INFORMATION POINT:

Regression analysis

In its simplest form regression analysis involves finding the best straight line relationship to explain how the variation in an outcome (or dependent) variable, Y, depends on the variation in a predictor (or independent or explanatory) variable, X. Once the relationship has been estimated we will be able to use the equation:

$$Y = b_0 + b_1X$$

in order to predict the value of the outcome variable for different values of the explanatory variable. Hence, for example, if age is a predictor for the outcome of treatment, then the regression equation would enable us to predict the outcome of treatment for a person of a particular age. Of course this is only useful if most of the variation in the outcome variable is explained by the variation in the explanatory variable.

In many situations the outcome will depend on more than one explanatory variable. This leads to multiple regression, in which the dependent variable is predicted by a linear combination of the possible explanatory variables. For example, it is known that the male peak expiratory flow rate (PEFR) depends on both age and height, so that the regression equation will be:

$$\text{PEFR} = b_0 + b_1 \times \text{age} + b_2 \times \text{height},$$

where the values b_0, b_1, b_2 are called the regression coefficients and are estimated from the study data by a mathematical process called least squares, explained by Altman (1991). If we want to predict the PEFR for a male of a particular age and height we can use this equation directly.

Often there will be many possible explanatory variables in the data set and, by using a stepwise regression process, the explanatory variables can be considered one at a time. The one that explains most variation in the dependent variable will be added to the model at each step. The process will stop when the addition of an extra variable will make no significant improvement in the amount of variation explained.

The amount of variation in the dependent variable that is accounted for by variation in the predictor variables is measured by the value of the coefficient of determination, often called R^2 adjusted. The closer this is to 1 the better, because if R^2 adjusted is 1 then the regression model is accounting for all the variation in the outcome variable. This is discussed, together with assumptions made in regression analysis, both by Altman (1991) and Campbell & Machin (1993).

In the preceding paper the outcome variable is ISQ-SR-N score and several independent variables were considered in the stepwise regression, which selected four for inclusion in the final model. Although this is the best model it still only accounts for 15.2% of the variation in ISQ-SR-N, because the R^2 adjusted is only 0.152. In other words, although the model explains a statistically significant amount of the variation, it still leaves most of it unexplained.

Further reading

Altman D.G. (1991) *Practical Statistics for Medical Research*. Chapman & Hall, London.

Campbell M.J. & Machin D. (1993) *Medical Statistics a Commonsense Approach*. 2nd edn. Wiley, London.

NICOLA CRICHTON