Statistical models – an overview

Ranstam J, Cook JA

A simple statistical model consists of an outcome and a sole explanatory (or predictor) variable. This type of model is sometimes described as bivariable as it includes only two variables. It is often used to estimate an “unadjusted” or “crude” effect that is, the influence of a single factor upon the outcome of interest without taking account of other factors which may also influence the outcome. To produce an “adjusted” estimate, other explanatory variables can be incorporated into the model (multivariable model). Different models can be used for various types of outcomes (e.g. logistic regression for a binary outcome) though all make assumptions regarding the relationship between the variables in the model in order to estimate their effects. Less commonly it may be or useful or necessary to model two or more outcomes together by assuming a joint (or multivariate) probability distribution and constructing a model accordingly (multivariate model) (1).

Multivariable models have two major uses: to explain or to predict. The first purpose is to define (with allowance for multiple factors) parameter estimates. In an observational study, an estimate of the factor of interest with adjustment for other important factors is often carried out this way. It may not fully resolve the issue due to unobserved factors and lead to a biased (though hopefully less) estimate. In a randomised trial, the reason for adjusting is mainly to reduce uncertainty of the estimates (increase the precision) by adjusting for strongly prognostic factors and accounting for randomisation (stratification) factors. Adjustment for chance baseline imbalance may also be done. Common to all study designs, the use of a multivariate model enables more individualized estimates (e.g. male and female specific) to be calculated.

The second purpose is to predict the outcome typically for an individual e.g. recurrence of disease following treatment (or classify individual properties). Development of a multivariable model in order to predict (predictive or prognostic model) differs from those above. The best prediction model is simply the model that predicts best(2). Predictive performance can be assessed in different ways. We will discuss the principles for model choice in each of these situations in more details later.

References

1. Hidalgo B, Goodman M. Multivariate or multivariable regression? Am J Public Health 2013;103:1-3.

2. Shmueli G. To Explain or to Predict? Statistical Science 2010;25:289–310.

Word count 355