

Selection of covariates for multivariable regression in observational studies: A review

Abhik Das

Statistics and Epidemiology Unit, RTI International, Rockville, MD, USA

Selection of covariates for multiple regression is a common problem in data analyses for public health studies. The issue is more acute for observational studies, whether cross-sectional or longitudinal, because (a) here, unlike clinical trials, we cannot rely on the study design or randomization scheme to correct for baseline imbalances among comparison groups, and (b) usually data are collected on numerous factors, all of which are potential candidates for inclusion in the regression model. Regardless of the specific nature of this model (linear for continuous outcomes, logistic for binary outcomes, hierarchical model for longitudinal outcomes, etc.) there are usually a large collection of covariates that can potentially be included. Consistency has been elusive in the public health research literature in dealing with this problem; indeed different papers from the same study have sometimes adopted slightly different strategies to cope with this issue. This review strives to develop some guidelines that can help public health investigators and statisticians adopt a more coherent approach towards this problem. Factors to consider in covariate selection include the available sample size, study design and analysis objectives, and specific quirks of the data structure. Automated variable selection algorithms, though frequently used, have substantial methodologic problems, and should only be used in a circumscribed manner with full awareness of their limitations. Ultimately, substantive considerations and the specific analytic objectives and approach should determine what strategy is used in each situation. Thus, the recommendations presented here should only be considered as guiding principles and not directives that fit every conceivable situation encountered by the researcher.