The Problem with Univariable Selection in Regression Modelling — and What To Do Instead

Theresa Ullmann (Medical University of Vienna, Austria), Georg Heinze (Medical University of Vienna, Austria), Franziska Kappenberg (Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany), Marc Henrion (Malawi-Liverpool-Wellcome Trust Clinical Research Programme, Blantyre, Malawi), Daniela Dunkler (Medical University of Vienna, Austria), on behalf of TG2 of the STRATOS initiative

Univariable selection, the screening of variables individually for statistical significance before inclusion in multivariable regression models, is one of the most widely used variable selection strategies in medical research. However, this approach lacks a theoretic justification, since a variable's association with an outcome in isolation does not reflect its role in a multivariable model. As a result, univariable selection can exclude important variables, retain spurious ones, and lead to biased estimates, invalid confidence intervals, and poor predictive performance.

Three key mechanisms explain this failure. *Masking* occurs when correlated variables obscure each other's true effects, causing important variables to appear non-significant in univariable analyses. *Amplification* occurs when a variable appears significant in a univariable model only because it is correlated with another variable that has a true effect. Finally, even when variables are uncorrelated, univariable analyses may still fail to detect important variables due to a *loss of precision*: including multiple relevant variables in a multivariable model reduces the residual error and can reveal predictors that univariable analyses miss.

We illustrate these problems using simulations and two real-world medical datasets (a breast cancer biomarker study and a study of lung health in schoolchildren). Combining univariable screening with subsequent backward elimination does not remedy the problem.

We discuss alternatives to univariable selection and offer concrete recommendations for variable selection tailored to descriptive, predictive, and causal modelling goals.