

Vine regression

Roger Cooke, Resources for the Future

Harry Joe, University of British Columbia, Canada

Bo Chang, University of British Columbia, Canada

Regular vines or vine copula provide a rich class of multivariate densities with arbitrary one dimensional margins and Gaussian or non-Gaussian dependence structures. The density enables calculation of all conditional distributions, in particular, regression functions for any subset of variables conditional on any disjoint set of variables can be computed, either analytically or by simulation. Regular vines can be used to fit or smooth non-discrete multivariate data. The epicycles of regression - including/excluding covariates, interactions, higher order terms, multi collinearity, model fit, transformations, heteroscedasticity, bias, convergence, efficiency - are dispelled, and only the question of finding an adequate vine copula remains. This article illustrates vine regression with a data set from the National Longitudinal study of Youth relating breastfeeding to IQ. Based on the Gaussian C-Vine, the expected effects of breastfeeding on IQ depend on IQ, on the baseline level of breastfeeding, on the duration of additional breastfeeding and on the values of other covariates. A child given 2 weeks breastfeeding can expect to increase his/her IQ by 1.5 to 2 IQ points by adding 10 weeks of Breastfeeding, depending on values of other covariates. Averaged over the NLSY data, 10 weeks additional breast feeding yields an expected gain in IQ of 0.726 IQ points. Such differentiated predictions cannot be obtained by regression models which are linear in the covariates.