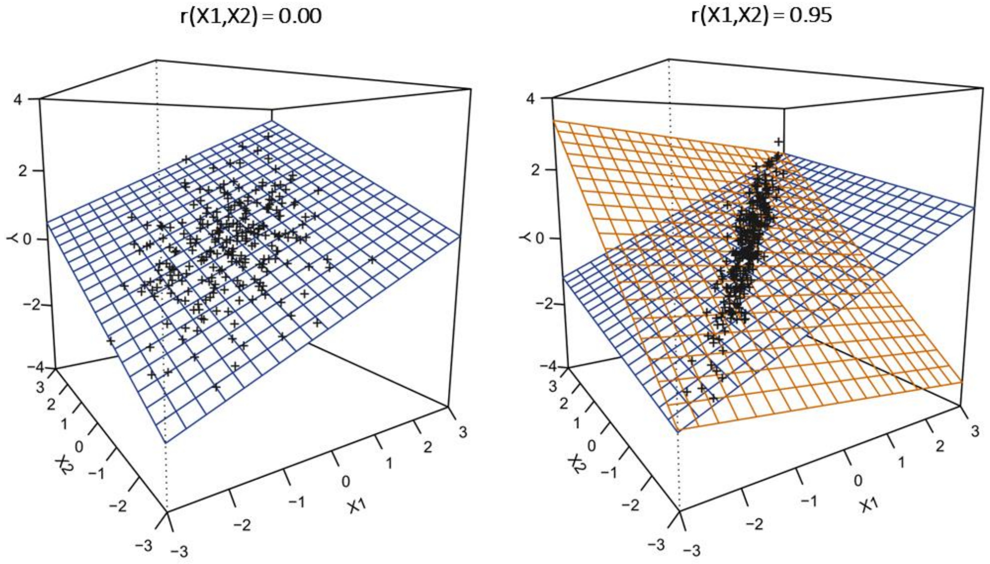
STATISTICAL SUPPORT
SWISS CENTER FOR AFFECTIVE SCIENCES

Multicollinearity

E-mail distributed on 09-08-2022

Dear all,

Today's communication is a reminder about multicollinearity. Multicollinearity occurs when the “correlation” between two or more effects in a model is problematically high (e.g., > 0.9). This means that the effects represent virtually the same information, and therefore estimating unique parameters for each is impossible.


r(X1,X2) = 0.00
r(X1,X2) = 0.95

The attached graph shows how to understand multicollinearity visually. On the left panel, predictors X1 and X2 are uncorrelated. Hence the cloud of points is evenly spread and the regression plane is well-defined. On the right panel, X1 and X2 are extremely highly correlated, such that the cloud of points is nearly a single line. However, a plane cannot be properly defined with only one line, and hence there is massive uncertainty in the regression coefficients. The orange and blue plane would both be consistent with the cloud of points. In statistical terminology, the **variance** in the second model is **inflated**, which will be reflected in large standard errors for the regression coefficients.

Multicollinearity can be diagnosed with **variance inflation factors (VIF)**, which in SPSS are available under "collinearity statistics", and in R through the `vif` function of package `car`. The conventional rule of thumb is that effects with VIF > 10 should be dropped from the final model. However, interaction effects may naturally exceed that limit, since by definition they are somewhat collinear with their corresponding main effects. For this reason, the `car::vif` function returns a generalized variance inflation factor (GVIF) that corrects for this bias and also handles multi-parameter effects.

Recall from my workshop on non-parametric data analysis (slide 22) that multicollinearity should generally be the first assumption to diagnose in your regression model, since this problem tends to distort parameter estimates and hypothesis tests the most severely.

Best,
Ben

PS: Note that normally it is good practice to keep correlated predictors in the same model, since this will allow you to isolate the unique effect of one variable when controlling for the others. It's only when the correlation becomes too high that this is no longer possible.

--
Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79