**Scale numerical DoE variables by standardization**

We have 5 DoE (Design of Experiment) parameters. While the 2 categorical variables are treated as dummy variables in our linear regression models, the 3 numerical variables, namely Adhesive.Out.Time, Prep..to.Bond.Time, and Contamination.Amount are pretreated so that they are standardized and balanced.

For each observed value of variable x, the standardized value

where is the mean of x, and the standard deviation.

Standardization is a common trick of data transformation that is commonly used in applied regression and clustering analysis. Subtracting the mean typically improves the interpretation of main effects in the presence of interactions, and dividing by the standard deviation puts all predictors on a common scale. [1]

Furthermore, while interaction terms can provide extremely important information about the relationship between the response and predictor variables, they also produce excessive amounts of multicollinearity, which is a problem because it can hide statistically significant terms, cause the coefficients to switch signs, and make it more difficult to specify the correct mode. [2] *Fortunately, standardizing the predictors is an easy way to reduce multicollinearity and the associated problems that are caused by these higher-order terms.* [2] Therefore, our regression analysis will be more or less benefited from standardization due to the presence of potential interactions.

References:

[1] <http://www.stat.columbia.edu/~gelman/research/unpublished/standardizing.pdf>

[2] <http://blog.minitab.com/blog/adventures-in-statistics-2/when-is-it-crucial-to-standardize-the-variables-in-a-regression-model>

Other references

<http://pj.freefaculty.org/guides/stat/Regression/StandardizedBeta/Standardized-1-lecture.pdf>

**Adding interactions to linear regression model**

Adding interaction terms to a regression model can greatly expand understanding of the relationships among the variables in the model and allows more hypotheses to be tested.

To illustrate, a simple linear regression (additive) model between a response variable y and two predictors and would be

while, in contrast, a model with interactions between and would take the form

In relationship to our regression analysis, the components and in a interaction term might be measurements or dummy variables in any combination. Interactions involving a dummy variable multiplied by a measurement variable are termed “slope dummy variables”. [1] For example, in our case, interaction between Contamination.Amount and a dummy variable representing one group of Contaminant.Type is a slope dummy variable, meaning the regression coefficient associated with it only contributes when the dummy variable equals 1.

When measurement variables are employed in interactions, it is often desirable to work with centering which makes the main effects in interaction models more interpretable. The coefficient a in the equation above, for example, represents the effect of when equals zero. Centering can also reduce problems with multicollinearity. [1] This again emphasizes the need for standardization of numerical DoE variables.

In our analysis, a linear regression model with up to 3-way interactions among all variables yields an adjusted-R2 of 96%, comparing to that of ?% from a simple additive linear regression model.

Reference:

[1] <https://en.wikipedia.org/wiki/Interaction_(statistics)>