troduction

- In statistics, the occurrence of several variables in a multiple regression model are **closely correlated** to one another, and carrying the same information, more or less. Multi-collinearity can cause strange results when attempting to study how well individual independent variables contribute to an understanding of the dependent variable, often undermining the analysis.
- In many analysis tasks, the variables under consideration are measured on different scales or levels. This would clearly distort any clustering analysis results. We can resolve this problem by ***standardizing*** the data prior to the analysis.
- Different standardization methods are available, such as the simple ***z standardization***, which re-scales each variable to have a mean of 0 and a standard deviation of 1.
- In most situations, however, ***standardization by range***(e.g., to a range of 0 to 1 or -1 to 1) is preferable. We recommend standardizing the data in general, even though this procedure can potentially reduce or inflate the variables influence on the clustering solution.

Transform Values

The following alternatives are available for transforming values:

- Z scores. Values are standardized to z scores, with a mean of 0 and a standard deviation of 1.
- Range -1 to 1. Each value for the item being standardized is divided by the range of the values.
- Range 0 to 1. The procedure subtracts the minimum value from each item being standardized and then divides by the range.
- Maximum magnitude of 1. The procedure divides each value for the item being standardized by the maximum of the values.
- Mean of 1. The procedure divides each value for the item being standardized by the mean of the values.
- Standard deviation of 1. The procedure divides each value for the variable or case being standardized by the standard deviation of the values.

Standardizing the Variables

- If variables are measured on different scales, variables with large values contribute more to the distance measure than variables with small values. In this example, both variables are measured on the same scale, so thats not much of a problem, assuming the judges use the scales similarly.
- But if you were looking at the distance between two people based on their IQs and incomes in dollars, you would probably find that the differences in incomes would dominate any distance measures. (A difference of only $100 when squared becomes 10,000, while a difference of 30 IQ points would be only 900. Id go for the IQ points over the dollars!).

- Variables that are measured in large numbers will contribute to the distance more than variables recorded in smaller numbers.
- In the hierarchical clustering procedure in SPSS, you can standardize variables in different ways. You can compute standardized scores or divide by just the standard deviation, range, mean, or maximum.
- This results in all variables contributing more equally to the distance measurement. Thats not necessarily always the best strategy, since variability of a measure can provide useful information.