Manual: 7.1.3. Selecting Independent Variables

Going back to the ideal gas law, we see that we have three independent variables pressure P, volume V, and temperature T in order compute the amount of gas n. This is what the model requires. Suppose, for a moment, that we did not know this.

The dataset will have many variables available to us. It may have various pressures and temperatures. There may also be vibrations, flows, levels, power consumptions and so on. From this multitude of offerings, we must select the three independent variables that will give us the information we need, i.e. the information needed to compute the dependent variable of interest.

There are two ways to do this selection. First, we may do it by using human ingenuity and understanding. This is the insight that came up with the ideal gas law. This requires such understanding and the time to input it all. As the understanding will not be definite in all cases but rather will be hypothesis driven, it will result in several loops of trial-and-error before a good model is found. This is a resource-consuming process.

Second, we may get the computer to do it for us based on data analysis. It is a major feature of the software that the selection of independent variables is done automatically with a very high chance of selecting a variable set that will yield a good model.

In the edit page for a dynamic limit under the heading of selecting independent variables, you will find the button for selecting the independent variables automatically. This function will assume that you have already defined the training times and exclude conditions. It will then get the data satisfying all these conditions and perform a correlation analysis between the dependent variable and all other variables as well as between all the other variables. It will select a set of variables that has high correlation with the dependent variable but little correlation with each other. That will ensure that we have enough but little duplicate information in the dataset as we want to keep the number of independent variables as low as we can without sacrificing too much accuracy. After the automatic selection, you may edit the list of selected variables if you like.

Having selected the independent variables, the model is ready to be trained.