

Approaches to Selecting Models

A common manual approach for selecting a model is to start with all the variables in the data set and eliminate the least significant variables, one variable at a time, based on p-values. After removing your first variable, you rerun the model and then remove the least significant terms from this model. You keep performing these steps until you're left with a model where only significant terms remain.

With a small number of predictor variables, a manual approach isn't too difficult. But with a large number of predictor variables, it's very tedious. You would need to continue this process until only terms with p-values lower than some threshold value, such as 0.05 or 0.10, remain.

Fortunately, SAS automates the model selection process for you. You specify the model selection approach to use, and SAS finds good candidate models. One approach is to use the all-possible regressions method, where all possible models are computed and ranked by the results. Then, to evaluate the models, you compare statistics side by side. These statistics include the familiar R-square and adjusted R-square values, along with a new statistic, the Cp statistic.

Another approach is to use stepwise selection techniques. Here you choose from the selection methods stepwise, forward, or backward. All three algorithms fit only a subset of all possible models to avoid the computation and time cost of fitting a massive number of models when the number of candidate predictors is large.

So, which automated approach is better? When you have a large number of potential predictors, the stepwise selection approach might be better. The all-possible regressions method produces more candidate models, but is more computationally expensive.

In either case, you'll need to look at some statistical measures, as well as your expertise to select a model. No one perfect model exists. One recommendation is to run several methods, look for commonalities across the results, and then narrow down your choice of models by using your subject-matter knowledge. You can then decide which model is best for you and your needs.