Print

Variable Selection

When we fit a multiple regression model, we use the p-value in the ANOVA table to determine whether the model, as a whole, is significant. A natural next question to ask is which predictors, among a larger set of all potential predictors, are important.

We could use the individual p-values and refit the model with only significant terms. But, remember that the p-values are adjusted for the other terms in the model. So, picking out the subset of significant predictors can be somewhat challenging. This task of identifying the best subset of predictors to include in the model, among all possible subsets of predictors, is referred to as variable selection.

In this video, we introduce some classical approaches to variable selection, but we'll revisit the problem in the module on predictive modeling. We also encourage you to explore these and other approaches to variable selection in the references provided.

One approach is to fit a full model and slowly remove terms one at a time, starting with the term with the highest p-value. This is referred to as backward selection. Let's look at an example. For the Impurity data, we fit a full model with two-way interactions. Reaction Time has the highest p-value. However, the caret next to the p-value indicates that Reaction Time is involved in interactions in the model, so we leave it in the model. The next highest p-value is Temp*Reactor. We remove this term and see that all the statistical output has updated, including the fit statistics and p-values for the model terms.

The next candidate for removal is Temp*Catalyst Conc. We remove this interaction term and continue this process until only terms with p-values below a chosen threshold remain. Here we stop removing terms when we hit the p-value threshold of 0.05. Stopping at a p-value of 0.05 is called a stopping rule, and the decision to stop at 0.05 was arbitrary. Typical stopping rules for explanatory modeling are p-value thresholds of 0.05 and 0.10. Note that there are other stopping rules we might consider. For example, we might stop at the model that has the highest RSquare Adjusted or the lowest Root Mean Square Error.

Later, we introduce two other important statistics for model selection: Akaike information criterion (or AIC) and the Bayesian information criterion (or BIC).

Returning to our example, here's our final reduced model. Let's compare the fit for the full model against this new reduced model. Both RSquare Adjusted and Root Mean Square Error are improved over the full model. So, this reduced model is better.

An alternative to backward selection is forward selection. With forward selection, instead of starting with a full model, we start with a model containing only the intercept. Then we slowly add terms to the model, one at a time, starting with the predictor with the lowest p-value. This continues until all the remaining terms that are not included in the model are above a specified p-value threshold. Or, we could use a different stopping rule to determine when to stop adding terms to the model.

A third classic variable selection approach is mixed selection. This is a combination of forward selection (for adding significant terms) and backward selection (for removing nonsignificant terms). As in forward selection, we start with only the intercept and add the most significant term to the model. We continue to add the most significant variables, one at a time. We use a p-value threshold to determine when to stop adding terms to the model. For example, we might set the p-value to enter the model at 0.05 or 0.10.

At each step, we look at the p-values for the terms in the model and compare the p-values to the threshold for removal. For example, we might set a p-value to leave the model at 0.10 or 0.15. If a p-value is greater than the threshold, the term is removed from the model. The mixed approach addresses a fundamental drawback of forward selection: terms might become insignificant after other terms have been added to the model. Mixed selection allows nonsignificant terms to be removed.

In this example, mixed selection results in a smaller model than backward selection. But, if we compare these two models, RSquare Adjusted is lower and Root Mean Square Error is higher for the mixed model. So, the backward selection model outperforms the mixed selection model.

A final approach that we'll only briefly introduce in this lesson is best subsets regression, or all possible models. Here, we fit all possible models from the combinations of the potential predictors. For example, consider the Impurity data with only the three continuous predictors: Temp, Catalyst Conc, and Reaction Time. How many possible models can we fit using these three predictors? We can fit one model with no predictors. This is the mean, or null, model. We can fit three models with only one predictor each. That is, we can fit one model with just Temp, one model with just Catalyst Conc, and one model with just Reaction Time. We can fit three models with two predictors each. And we can fit one model with all three predictors. This gives us eight potential models. In best subsets regression, we fit all these models and then compare the models to pick the one "best" model.

Note that best subsets regression can quickly get out of hand as we increase the number of potential predictors. For example, if there are 10 potential predictors, then there are two to the tenth or 1024 potential models.

In the next video we see how to use the Effect Summary table to do variable selection. For more information on Stepwise selection and All Possible Models, see the Read About It for this module.

Close