Print

Interpreting p-Values and Parameter Estimates

You should be cautious when you're interpreting and reporting statistical quantities that are produced by these methods. Using automated model selection results in biases in parameter estimates, predictions, and standard errors, incorrect calculation of degrees of freedom, and p-values that tend to err on the side of overestimating significance. Some of these problems result from using p-values to select variables. The p-values are intended for testing one hypothesis, not dozens of hypotheses in many models with overlapping sets of predictors. Using other criteria to determine which variables enter and leave the model might alleviate some of these problems. In the next section, we'll discuss using information criteria and adjusted R-square for variable selection.

Another cause of the biases in parameter estimates and p-values originates from using the same sample both to choose the model and to assess it. This can be alleviated by splitting the sample into one portion for model development, and a holdout portion for model assessment. For example, model selection could be performed on half of the available data. Then the model could be applied to the holdout data set, and parameters and p-values from this holdout could be reported. You're assessing how well your model performs on a different sample of data than the one you used to develop the model.

You want to avoid overfitting the model on a single sample, and ensure that the model will generalize to other data sets. Often researchers, especially those analyzing designed experiments, will not have enough data to partition it into training and holdout sets. With small or moderate data sets, data splitting is inefficient. The reduced sample size can severely degrade the fit of the model.

However, computer-intensive methods, such as k-fold cross validation and bootstrap methods, were developed so that all the data can be used for both fitting and honest assessment. In k-fold cross validation, you train and assess the model on k total different partitions of the data for the same model. The results from each holdout set can then be averaged to interpret how well the model generalizes to new data. Alternatively, bootstrapping is a resampling method that tries to approximate the distribution of the parameter estimates in order to obtain correct standard errors and p-values.

Close