

Contents

1 Theoretical Aspects of Fitting Models 2

1.1 The Law of Parsimony 2

1.2 Model building 2

1.3 Overfitting 2

1.4 Variable-Selection Procedures 2

1.5 Validation and Testing 2

1 Theoretical Aspects of Fitting Models

1.1 The Law of Parsimony

Ockham's razor, sometimes known as the law of parsimony, is simply a maxim that states that simple explanations are usually better than complicated ones. **Ockham's razor** was originally proposed by a monk named William of Ockham. (He did not call it "Ockham's razor" or even "my razor." This is a name that has been given to it over time.)

Another version of this principle is the Law of parsimony. This says that if you are choosing between two theories, choose the one with the fewest assumptions. Assumptions here means claims of fact that have no evidence. A theory that doesn't have many assumptions, and is very simple, is called a parsimonious theory.

In the context of statistics, the law of parsimony can be interpreted as an adequate model which requires the fewest independent variables is the preferred model.

1.2 Model building

The traditional approach to statistical model building is to find the most parsimonious model that still explains the data. The more variables included in a model (overfitting), the more likely it becomes mathematically unstable, the greater the estimated standard errors become, and the more dependent the model becomes on the observed data. Choosing the most adequate and minimal number of explanatory variables helps to find out the main sources of influence on the response variable, and increases the predictive ability of the model. As a rule of thumb, there should be more than 10 observations for each variable in the model.

1.3 Overfitting

Overfitting occurs when a statistical model does not adequately describe of the underlying relationship between variables in a regression model. When overfitting happens, the model predicts the fitted data very well, but predicts future observations poorly.

Overfitting generally occurs when the model is excessively complex, such as having too many parameters (i.e. predictor variables) relative to the number of observations. A model which has been overfit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data.

1.4 Variable-Selection Procedures

In regression analysis, variable-selection procedures are aimed at selecting a reduced set of the independent variables - the ones providing the best fit to the model, in keeping with the Law of Parsimony.

1.5 Validation and Testing

When you have sufficient data, you can subdivide your data into three parts called the training, validation, and test data. Rather than estimating parameter values from the entire data set, the data set is broken into three distinct parts. During the *variable selection* process, models are fit on the training data, and the prediction error for the models so obtained is found by

using the validation data. Validation is the process of using part of a data set to estimate model parameters, and using the other part to assess the predictive ability of the model. Validation can be used to assess whether or not overfitting has occurred.

This prediction error on the validation data can be used to decide when to terminate the selection process or to decide what effects to include as the variable selection process proceeds. Finally, once a selected model has been obtained, the test set can be used to assess how the selected model generalizes on data that played no role in selecting the model.

- 1 The training set is the part that estimates model parameters.
- 2 The validation set is the part that assesses or validates the predictive ability of the model.
- 3 The test set is a final, independent assessment of the models predictive ability.

A validation set is a portion of a data set used to assess the performance of prediction or classification models that have been fit on a separate portion of the same data set (the training set). Typically both the training and validation set are randomly selected, and the validation set is used as a more objective measure of the performance of various models that have been fit to the the training data (and whose performance with the training set is therefore not likely to be a good guide to their performance with data that they were not fit to).

It is difficult to give a general rule on how many observations you should assign to each role. One important textbook recommended that a typical split might be 50% for training and 25% each for validation and testing.