

StatNews #19: Tree-Based Regression Methods

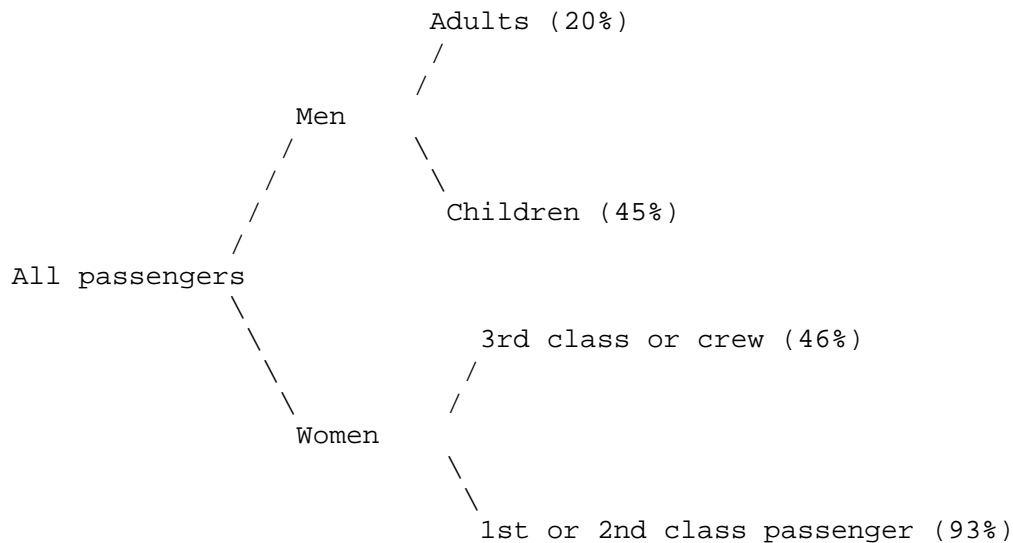
March 11, 1997

For many studies, addressing the primary research question requires uncovering and understanding the relationships among several factors. The standard tools for this purpose are various forms of general or generalized linear models (e.g., linear or logistic regression, analysis of variance). This newsletter discusses an alternative approach, "tree-based" methods, which are inherently more flexible, can easily handle complicated interactions among factors and large numbers of factors, and give results that are simple to interpret.

Tree-based methods involve dividing the observations into groups that differ with respect to the variable of interest. For example, suppose we want to know which passengers on the Titanic were most likely to survive the ship's sinking, and what characteristics were associated with survival. In this case, the variable of interest is survival. We could divide the passengers into groups based on age, sex, and class, and look at the proportion surviving in each group. A tree-based procedure automatically chooses the grouping that results in homogeneous groups that have the largest difference in proportion surviving.

In the Titanic example, the tree-based method first divided the observations into men and women. The next step is to subdivide each of the groups based on another characteristic. Men were divided into adults and children, while women were divided into groups based on class. Notice that the process of subdividing is separate for each of the groups. This is an elegant way of handling interactions that can become complicated in traditional linear models.

When the process of subdivision is complete, the result is a classification rule that can be viewed as a tree. For each of the subdivisions, the proportion surviving can be used to predict survival for members of that group. The structure of the tree gives insight into which characteristics of the passengers are related to survival. A tree for Titanic survival, with the proportion surviving in each subgroup, is given below.



Tree-based methods have several attractive properties when compared to traditional methods. They provide a simple rule for classification or prediction of observations, they handle interactions among variables in a straightforward way, they can easily handle a large number of predictor variables, and they do not require assumptions about the distribution of the data. However, tree-based methods do not conform to the usual hypothesis testing framework.

There are several tree-based methods that differ with respect to the types of variables allowed, the way groups are chosen, and the way groups are split. The most common methods are Classification and Regression Trees (CART) and Chi-Square Automated Interaction Detection (CHAID). CART and similar methods allow the response and grouping variables to be either categorical or continuous. CART methods are implemented in SYSTAT version 7 and in S-Plus, and will supposedly be implemented in SAS within a year. CHAID and similar methods require the response variable to be categorical. CHAID methods are available in an SPSS add-on module and in the SAS macro %TREEDISC.

If you are interested in using a tree-based procedure, there are several issues to consider, including how to choose the right size tree and how to assess the performance of the tree. Feel free to contact our office for assistance.