Abstract
Regression and classification trees are methods for analyzing how a dependent variable is correlated with independent variables. Proc Hpsplit and Proc Dtree can both create decision trees that look similar. Both begin with a single node followed by increased number of leaves. However, they focus on a different purpose. This paper is a preliminary introduction differences between Proc Hpsplit and Proc Dtree.

Introduction

When we want to explore the relationship to the variables and outcome, how the effect of variables on the outcome, Proc Hpsplit is an useful tool. On the other hand, decision tree with proc dtree is to find out the most desired output given the combination of variables.

Outline of paper:

1: the difference between decision tree and dtree- some decision tree will not show the what's going on in the inner components. Proc dtree can make the decision and show the inner components directly.
2: I use popular data called Iris dataset to analyze how to split with proc hpsplit and proc dtree (theory and math)
3: Discussion of above

Sample graphics:

Proc Dtree

Obs	_STNAME_	_STTYPE_	_OUTCOM_	_SUCCES_
1	species	S	Setosa	petal_width
2			Versicolor	
3			Virginica	
4	petal_width	width	less_than_8	species
5			between_8and_16.5	
6			greater_than_16.5	
7	species	S	Setosa	petal_length
8			Versicolor	
9			Virginica	
10	petal_length	length	less_than_49.5	species
11			greater_than_49.5	
12	species	S	Setosa	
13			Versicolor	
14			Virginica	

Obs	_EVENT1	_PROB1	_EVENT2	_PROB2	_EVENT3	_PROB3
1	less_than_8	0.2	between_8and_16.5	0.6	greater_than_16.5	0.2
2	Setosa	0.5	Versicolor	0.3	Virginica	0.2
3	less_than_49.5	0.2	greater_than_49.5	0.8		.

Proc hpsplit

When we use proc hpsplit to get the 2 way splitting (but split number is less than number of levels ,we have 3 levels for Y -Species). This 2 way split is good, but maybe not desired splitting for this case.