DM STAT-1 CONSULTING **BRUCE RATNER, PhD**

574 Flanders Drive North Woodmere, NY 11581 br@dmstat1.com
516.791.3544 ~ fax 516.791.5075 www.dmstat1.com

Interpretation of Coefficient-free Models

Bruce Ratner, PhD

The ordinary regression model is the thought of reference when beginners-through-advanced model builders, and model end-users (e.g., those who know all about understanding and implementing models, but cannot build one) hear the words “new kind of model.” Data analysts use the regression concept and its prominent characteristics when judiciously evaluating an alternative modeling technique. This is because the ordinary regression paradigm is the underpinning for the solution to the ubiquitous prediction problem. End-users with limited statistical background undoubtedly draw on their educated notions of the regression model before accepting a new technique. Model builders go back to their first steps of the statistical model-building paradigm. New modeling techniques are evaluated by the coefficients they produce. The coefficients are deemed essential because they are always used as a measure of *rank-order importance* of variables that drive a model, i.e., that put variables into their proper places in relation to each other in predicting and explaining a model. If the new coefficients impart comparable information to the prominent characteristic of the regression model – the regression coefficient – then the new technique passes the first line of acceptance. If not, the technique is summarily rejected. A quandary arises when a new modeling technique, like machine learning methods, produces models with no coefficients.

I present two machine learning models, one pretty popular, and one not-yet popular. [1] The *pretty* one does not trouble data analysts about the *missing coefficients* because nearly anyone can understand the highly accepted model, regardless of one's statistical background. Yet, data analysts are not as forgiving with the not-yet popular one. The latter model requires a passing introductory training in machine learning, which data analysts typically have no such exposure. Accordingly, they virtually always shout "Where are the coefficients?"

Consider the pretty one, but hardly ever thought of as a machine learning method – CHAID. See Response CHAID Tree in Figure 1, below. The CHAID Tree can *loosely* be interpreted: The overall Response of 10% (from a population of size 1000) is explained and predicted by primarily Martial Status, and secondarily Gender and Pet Ownership.

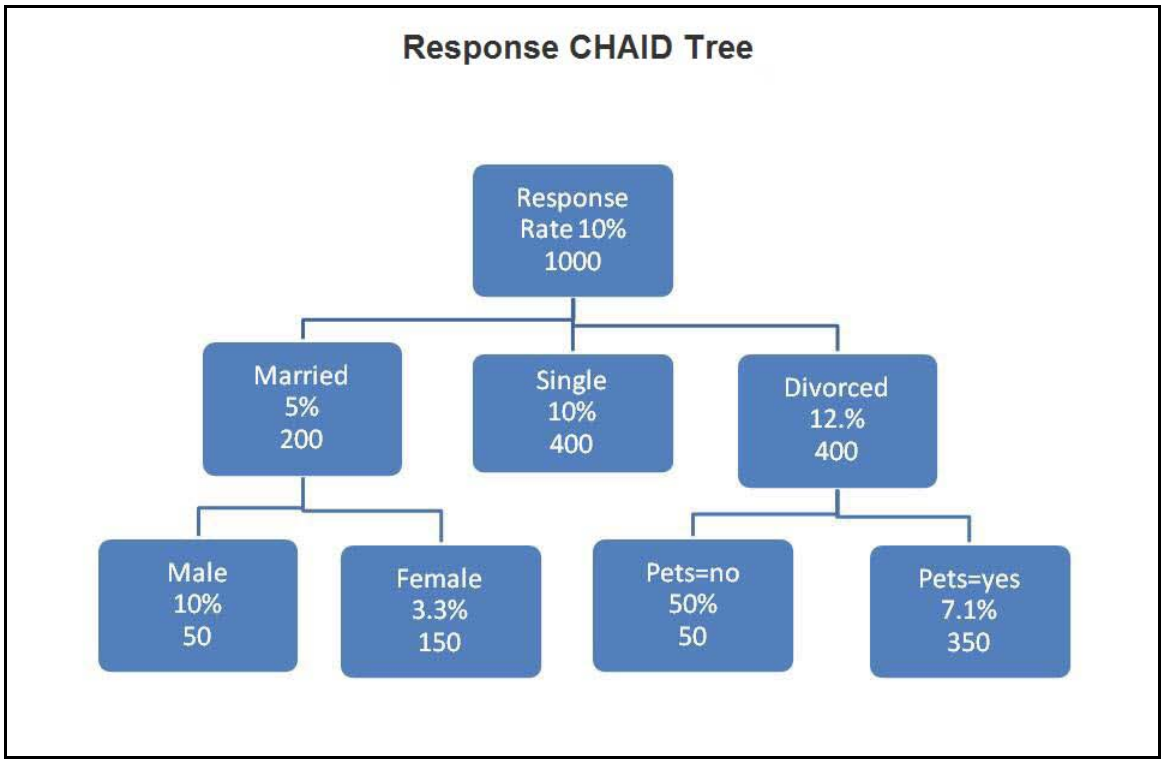

Figure 1. Response CHAID Tree

In a CHAID model, the rank-order importance of variables is determined by which predictor variable splits first; this variable is arguably statistically declared the most predictive. The rank-order of the other variables, actually, combination (interaction) variables, are dependent on a three-way trade-off analysis among 1) number of times a variable appears in two-, three-, and four-interactions segments (five-way and greater interactions are rarely seen), 2) the sizes of interaction segments, and 3) the magnitudes of the target variable (Response is this example) in the interaction segments. [2]

Now for the not-yet popular GenIQ Model that consists of a two-part output, a *parse* tree, and the computer code that cryptically *explains* the tree. [3] Observe the Response GenIQ Model in Figure 2, below. Unlike in CHAID, the rank-order importance of variables in a GenIQ Model is not concisely described. Although not difficult to understand, the GenIQ Model rank-order importance measure requires much space, not available here. [4]

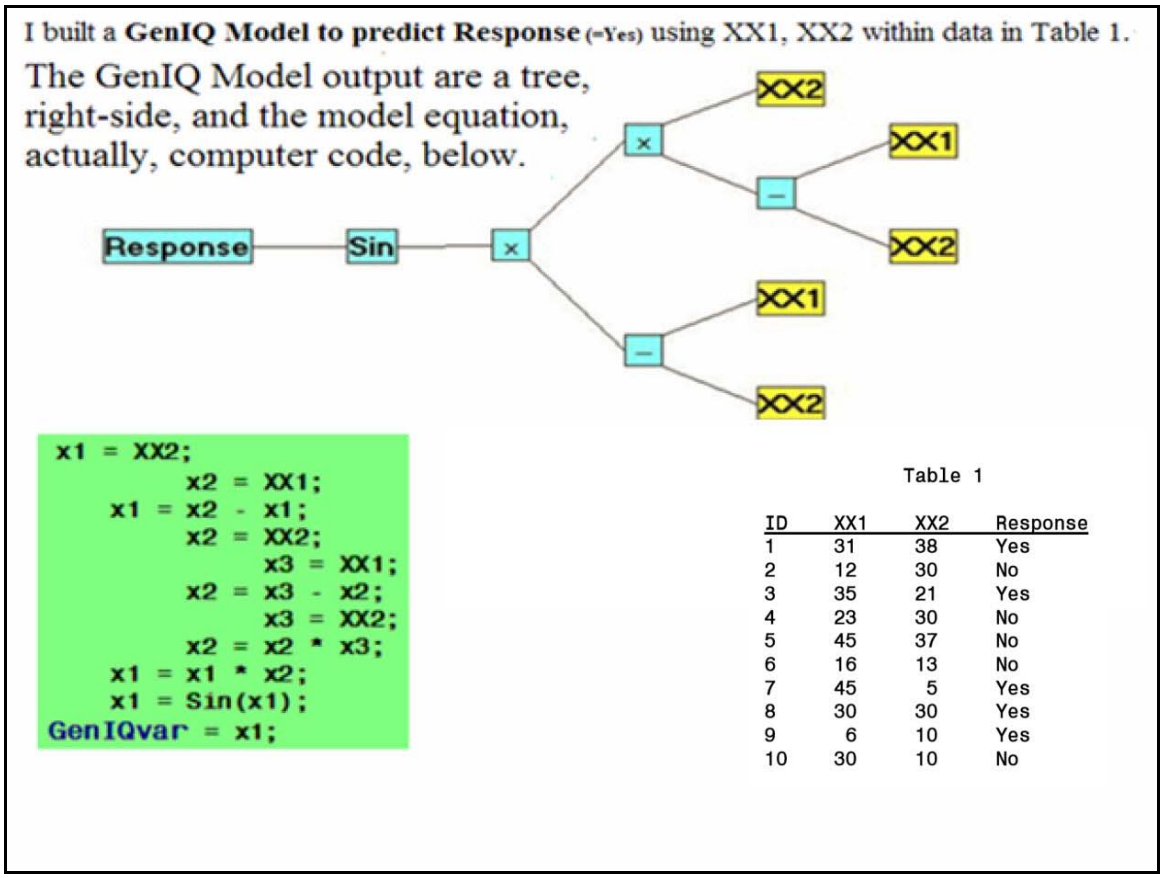

Figure 2. Response GenIQ Model

The primary purpose of this article is to present a method for calculating a *quasi-regression coefficient*, which provides a frame of reference for evaluating and using coefficient-free models. Secondarily, the quasi-regression coefficient serves as a *trusty assumption-free alternative* to the regression coefficient, which is based on an implicit and hardly-tested assumption necessary for reliable interpretation. [5]

References:

1 - Not-yet Popular Model

2 - Personal Observation of Size 1: I know CHAID is widely used as an end-result model (as opposed to using CHAID as a data-mining model), but a model based on only one *main effect* predictor variable, and many, many two-, three-, and four-way interactions, which are identifying very small (i.e., unreliable) segments of the population, is not for my taste. Please keep in mind that I am not discounting CHAID. See my nine uses of CHAID beyond its original intent. [1a]

3 - A parse tree is a visual representation of a computer program, or a model of the type GenIQ Model produces.

4 - What is the GenIQ Model?

5 - This article is also in my book *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Ming Big Data*, Chapter 17.