Sawtooth Software

RESEARCH PAPER SERIES

CBC/HB for Beginners

John Howell,
Sawtooth Software, Inc.

530 W. Fir St.
Sequim, WA 98382
(360) 681-2300
www.sawtoothsoftware.com

CBC/HB for Beginners

John Howell, Sawtooth Software, Inc.

March 2009

Introduction

CBC/HB has become the most popular method for estimating utilities from Choice Based Conjoint Studies based on recent research by Sawtooth Software. However the process of estimation has historically been hard to explain to a non-technical audience. There are few resources available for non-technical people to understand how CBC/HB works. This paper attempts to explain CBC/HB in very simple terms. It avoids most of the mathematics and technical terms that are commonly used to describe the technique and instead uses heuristics and examples. For additional information and implementation details see the additional resources section at the end of the paper.

Individual Level Models

The basic problem that CBC/HB is trying to solve is to create individual-level utilities for each respondent. A utility is a number representing the attractiveness of each feature in a conjoint study. Individual utilities are helpful because they allow for easy segmentation and they provide a way to detect different groups. Since people are different and have different preferences, it is not always accurate to roll a sample together into a single set of utilities.

To illustrate this, imagine you are studying a polarizing brand. Half the people you interview love the brand and the other half hate it. If you simply look at the average you would come to the conclusion that the sample is ambivalent toward the brand. That is probably the worst conclusion that could be drawn because it is wrong for the entire sample, leaving the brand preference unaddressed. This is an extreme example, but less extreme examples of this are commonly seen in actual studies.

If you have individual-level utilities, you can detect the segments that disagree and target them separately. It also makes it possible to predict individual market choices and build accurate what-if simulators that are sensitive to different preferences.

With CBC, however, there is a problem with directly calculating individual utilities. Each respondent only provides a small amount of information within a CBC interview. Typical CBC surveys often have the respondent make as few as 8-12 separate selections. It would seem difficult to try to estimate the preferences for thousands of different product combinations from the relatively small amount of information provided. But, CBC/HB provides a clever method, based on the rules of probability, to overcome the problem of sparse information.

CBC/HB Algorithm

Instead of estimating each respondent's utilities individually, the algorithm estimates how different the respondent's utilities are from the other respondents in the study. This is a much easier task than estimating each respondent's individual utilities independently. The algorithm estimates the average utilities for the entire sample and then uses the respondent's individual data to determine how each respondent differs from the sample averages. The algorithm will then adjust each respondent's utilities so that they reflect the optimal mix of the individual respondent choices and the sample averages.

The optimal mix of respondent data and sample averages is determined by the amount of data a respondent provides and the amount of variance in the sample average. The more variance in the sample averages, the more the algorithm will rely on the individual respondent's data. Conversely, the more choice tasks there are for each respondent, the less influential the sample averages will be.

Imagine you are interviewing U.S. high school students and one of the questions is about their age. In U.S. high schools, the average age is probably between 15 and 16 and nearly all students are between 13 and 19. If you were to interview a student who claimed to be 21 you would probably be suspicious and would discount the answer because it did not fall in the expected range. On the other hand if you were interviewing in a shopping mall, the average age may still be between 15 and 16, but interviewing a 21 year old would not seem unusual because the variability of ages is much larger. HB provides a formal way of incorporating context into our data collection.

A little side note about terminology: the sample averages represent information we have before we look at the individual respondent's data so it is called the prior. It is also sometimes called the "upper model" in technical papers. The final utilities we get after looking at the respondent's data are called the posterior utilities. The "lower model" refers to the estimation of the respondent's preferences.

There is a little problem generating the respondent utilities from the sample utilities since the sample utilities depend on what the individual utilities are. Every time the individual utilities are updated, the sample average would need to be updated. In turn every time the sample average is updated, the individual utilities would need to be updated. To overcome this problem CBC/HB goes though a series of iterations. It starts with an arbitrary, made-up sample average. (It uses a mean of 0 for all levels as a starting point.) It then estimates what the individual utility scores would be assuming the sample averages were actually at the starting point. After all the individual preference scores have been calculated, the algorithm updates the sample average and then repeats the process. In the beginning, the estimates are not very good since the sample average is not accurate, but after a few cycles the estimates begin to stabilize. Once the algorithm stabilizes, the sample averages stop changing much between iterations. The process is said to have converged. The cycle is then run a few thousand more times and the results from each iteration are saved. These values make up the saved iterations. The final results are calculated by taking the average of the saved sample averages and the average of the saved individual's utilities.

Some Mathematical Details

The CBC/HB algorithm is controlled by the rules of probability. There are two important probabilities that we are dealing with in the CBC/HB algorithm. The first is the probability that a respondent will select a specific concept in a choice task given a specific set of utilities. We will call this the likelihood. The second probability that we are concerned with is the probability that the respondent's utilities are consistent with the pattern of utilities observed in the rest of the respondents. We will call this the sample density.

One of the basic rules of probability is that the joint probability of two events is the probability of the first times the probability of the second provided that the two events do not depend on one another. This rule is used heavily in the CBC/HB algorithm. First we make an assumption that each choice task is evaluated independently. That means that we assume that respondents have "no memory" about previous choice tasks they answered. Another way of looking at this is that a respondent would have answered each choice task exactly the same way if the choices were presented in a different order. This means that for a specific set of utilities we can determine the probability of a respondent answering a choice exercise a certain way by multiplying together the probabilities for each individual task. The closer the estimated utilities are to the respondent's true utilities, the higher the total probability will be.

The second place we use this rule is to combine the sample density and the individual likelihood. We multiply the respondent's likelihood, the total probability from the respondent's choices assuming a specific set of utilities, by the sample density, the probability that those utilities came from the same population as the rest of the sample. This gives us the probability that a set of choice tasks were answered according to a given set of utilities and that those utilities came from the same population as the rest of the sample's utilities.

CBC/HB assumes that the respondent answers choice tasks according to a Multinomial Logit model (MNL). MNL considers the probability of the specific alternative being chosen related to the proportion of the total utility for that concept relative to the total utility for all the concepts. The formula is:

$$P(chosen) = \frac{e^{(U_{chosen})}}{e^{U_1} + e^{U_2} + e^{U_3} + \ldots}$$

In the formula, the U_{chosen} represents the total utility for the chosen concept; the $U_{1, 2, \ldots}$ are the total utilities for each concept including the chosen one. Notice that this is essentially a percentage calculation, since the denominator of the fraction includes the numerator plus the other non-chosen products. The total utility for each concept is calculated by adding the utility associated with each attribute level included in the task. The only way to change the probability then is by changing the individual utility levels.

The sample density is calculated based on a multivariate normal distribution. The mean of the multivariate normal distribution is the mean of all respondent utilities. The variance of the multivariate distribution is based on the variance of the sample's utilities. This means that unusual utilities have a very low probability while utilities that are close to the sample average will have a high probability. This does not mean that the individual utilities will follow a normal distribution. Since they are combined with the MNL portion, they can be nearly any shape.

The algorithm is trying to maximize individual likelihood times the probability from the multivariate normal distribution. This means that it may be necessary to "shrink" extreme utilities so they don't fit the MNL model quite as well. At the same time the probabilities from the multivariate normal distribution would be higher so the overall fit would be better. This shrinkage provides stability to the estimation and improves the prediction of the entire model. Unusual or unlikely respondents are down-weighted, while consistent, predictable respondents are not affected much by the shrinkage.

The amount that utilities are shrunk depends on the variance of the multivariate normal distribution. The larger the variance the more the probabilities flatten out so they have less effect on the total probability. On the other hand, if the variance of the normal distribution is small, the probabilities will quickly get very small as they move away from the mean. This results in extreme utilities being shrunk dramatically toward the mean. Because the variance depends heavily on the individual data, the algorithm is self-regulating. If the respondents come from diverse segments, they will have a large variance, but if the respondents are consistent they will have a small variance. The shrinkage provides just enough pull to stabilize the results.

This paper has been light on technical details and is not sufficient to reproduce CBC/HB. There are a few points in the algorithm that have been glossed over in the interest of simplicity, but are very important in the final algorithm. Please see the resources listed in the Additional Resources section for a more complete treatment of CBC/HB and related techniques.

Additional Resources

Sawtooth Software, Inc. (2007) *"2007 Customer Survey Results"* Sawtooth Solutions Spring 2007, Available from http://www.sawtoothsoftware.com

Sawtooth Software, Inc. (2005) "*CBC/HB System for Hierarchical Bayes Estimation Version 4.0 Technical Paper*" Technical Paper available at http://www.sawtoothsoftware.com.

Lenk, Peter, "Bayes Notes", http://webuser.bus.umich.edu/plenk/downloads.htm, (Accessed March 16, 2009)

Rossi, Peter E., Allenby, Greg M. and McCulloch, Robert (2005) "Bayesian Statistics and Marketing," John Wiley & Sons Ltd., West Sussex.

Gelman, A., Carlin, J. B., Stern H. S. and Rubin, D. B. (2004) "Bayesian Data Analysis," Chapman & Hall/CRC, Boca Raton.