Sawtooth Software

RESEARCH PAPER SERIES

Monotonicity Constraints in Choice-Based Conjoint with Hierarchical Bayes

Richard M. Johnson,
Sawtooth Software
2000

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

Monotonicity Constraints in Choice-Based Conjoint with Hierarchical Bayes

Richard M. Johnson
Sawtooth Software, Inc., April, 2000

Conjoint studies frequently include product attributes for which almost everyone would be expected to prefer one level to another. For example, most people would prefer high quality to low quality, long life to short life, and low cost to high cost. However, estimated part-worths sometimes turn out not to have those expected orders. This can be a problem, since price utilities with the wrong signs or slopes are likely to produce models yielding nonsense results. Perhaps even more important, anomalous part worths can undermine users' confidence in the results.

Several authors have argued that conjoint results can be improved by constraining part-worths to have specified signs or order relations. Others have argued that constraining levels of an attribute to have a certain order tends to inflate its importance. Thus, it is not entirely clear whether monotonicity constraints should be used in conjoint analysis.

Hierarchical Bayes (HB) methods have recently become available to marketing researchers, and there is ample evidence of their superiority for estimation of conjoint part-worths. However, commercially-available HB software does not provide any ability to enforce monotonicity constraints. Thus, it is appropriate to assess alternative methods of enforcing monotonicity constraints in the HB context, as well as the general advisability of doing so. This report presents a comparison of ways to enforce monotonicity constraints when using HB to estimate conjoint part-worths from choice data. Results are compared for several methods:

(1) Tying offending values of unconstrained point estimates following estimation
(2) Tying offending values of individual HB draws following estimation
(3) A method employing importance sampling
(4) A method employing log normal distributional assumptions
(5) A method employing "Tobit-like" transformations
(6) A simultaneous tying method, a combination of methods 2 and 5.

Methods are compared with respect to hit rates in prediction of individual choices for holdout concepts, as well as errors predicting aggregate holdout shares. The main conclusions are that

> Constraints usually provide improvements in hit rates, and sometimes provide improvements in share predictions.
>
> The "tying" methods appear most successful and are also easy to implement.

We describe each method and examine its performance using one data set, and then check the relative performance of methods with a second data set.

The TV Data

The first data set was provided by Huber, Orme, and Miller (1999), consisting of choices for 352 commercial respondents who considered TV sets using six attributes:

Brand (3 levels)
Screen Size (25, 26, 27 inch)
Sound Quality (mono, stereo, surround)
Channel Blockout ("CB", present/absent)
Picture-in-picture ("PIP", present/absent)
Price ($300, $350, $400, $450)

Price, screen size and sound quality seem reasonable candidates for monotonicity constraints (although some respondents could conceivably want smaller screens or less complicated electronics). Respondents first completed 18 customized choice tasks, each with 5 alternatives, without a "None" option. Each respondent also completed 9 holdout choice tasks, also consisting of 5 alternatives, each specified on 6 attributes. The 18 calibration choice tasks seen by each respondent were designed randomly and were unique. However, respondents in each of 4 random groups saw identical holdout tasks. Therefore, we are able to measure not only holdout hit rates, but also mean absolute errors (MAEs) for aggregate share predictions.

Share predictions were only made for the holdout tasks each respondent actually saw. The holdout tasks were deliberately designed to be difficult, often containing very similar and sometimes even duplicate alternatives (to offer a severe IIA challenge).

The Unconstrained Solution

The unconstrained solution was obtained using off-the-shelf CBC/HB (Sawtooth Software, 1999). There were 34,000 initial "burn-in" iterations, followed by 100,000 subsequent iterations during which each 10^{th} was saved. Point estimates of each respondent's part worths were obtained by averaging those 10,000 saved draws. This was a much larger number of draws than would normally be saved, but they were needed for the importance sampling method. Hit rates and MAEs for share predictions were computed similarly for each method, so those details are provided here.

> Hit rate was computed by simulating each respondent's choices for 9 holdout choice sets. If the respondent was predicted to choose an alternative which happened to be duplicated within that choice set, that prediction was scored as correct if he chose either of the duplicate items.
>
> Share predictions were made using the Randomized First Choice method described by Orme and Baker (2000). Each respondent's choices were simulated

1,000 times, adding iid random normal "attribute error" to point estimates of his part-worths for each simulation. Actual and predicted choices of duplicate alternatives were scored as half-choices for each of the duplicate items. The amount of random error was chosen empirically.

The unconstrained solution had a hit rate of **.655**, and MAE for share predictions of **3.22**.

Severity of Constraints

We tried constraining different combinations of attributes. The following table shows the number of monotonicity violations that would be corrected by constraining each attribute other than Brand:

Severity of Constraints

Attribute	Violations	Opportunities	%Violations
Screen	112	704	16
Sound	152	704	22
CB	86	352	24
PIP	47	352	13
Price	180	1056	17

Consider the first row of the table. The Screen size attribute had 3 levels, with 2 pairs of adjacent levels, so with 352 respondents there were 2 x 352 = 704 opportunities for a monotonicity violation. In the unconstrained solution there were actually 112 such violations, for a rate of 16%. CB and PIP each had two levels, so only one monotonicity constraint was involved in each. Price had 4 levels, so there were 3 pairs of adjacent levels. The percentages of violations were highest for Sound, and CB, and lowest for PIP.

We have predicted holdout choices using part-worths estimated by several different combinations of these constraints, including price alone (3 order constraints), price, screen size, and sound quality (7 order constraints), and the above plus requiring CB and PIP to be positive (9 order constraints). The best results were generally obtained when constraining price, screen size, and sound quality, and those are the results we shall report here.

Enforcing Constraints by Tying Values After Estimation

This method consists of doing HB estimation without any constraints, and then applying constraints afterwards by tying the values in each offending pair of point estimates of part-worths. Recursion is necessary because for an attribute with several levels, tying one pair of levels may cause an order violation involving some other level. The algorithm recycles through all pairs of levels, tying each offending pair, until no pair of levels violates a specified order relationship by more than some small amount.

This method was originally included only as a "straw man," with expectation that it would do relatively poorly. Here are the resulting hit rates and share prediction errors for three different severities of constraint. We include the unconstrained solution for reference.

Method	**Hit Rate**	**MAE**
Unconstrained	**.655**	**3.22**
Tying Post Hoc	**.659**	**3.22**

The table gives the proportion of individual choices correctly predicted, and the mean absolute error of predictions of choice shares. Although tying the offending values of the point estimates is the easiest conceivable way of dealing with constraints, it does quite well. The hit rate is slightly higher than the hit rate for the unconstrained solution, and the MAE values are identical to two decimals.

This tying method might be even more effective if applied to individual draws rather than to the point estimates. When tying is done at the level of individual draws, the resulting averages will have fewer ties, and might conceivably contain more information. That was tried next, with these results (we repeat previous results for reference):

Method	**Hit Rate**	**MAE**
Unconstrained	**.655**	**3.22**
Tying Post Hoc	**.659**	**3.22**
Tying Draws Post Hoc	**.664**	**3.31**

Imposing constraints by tying individual draws following estimation produces a higher hit rate than tying the point estimates, although share predictions are a little worse.

Enforcing Constraints by Importance Sampling

The next method was an importance sampling procedure proposed by Allenby (2000):

> 1) Estimate the model without imposing any constraints and save the betas for each individual. The average of these beta draws provides an estimate of the posterior mean without imposing restrictions.
>
> 2) To estimate the posterior mean of beta with restrictions, only use the betas that conform to the ordering. For example, to estimate the posterior mean of a person's price coefficient, calculate the mean using only those draws of the price coefficient that are negative.
>
> 3) Estimate the covariance matrix using all the draws, not just the draws that conform to the ordering.

This procedure was implemented by reading the 10,000 random draws for each respondent saved while estimating the unconstrained solution, and retaining only those draws conforming to the specified monotonicity constraints. This was tried for three severities of constraints. As would be expected, many fewer draws are retained when many constraints were imposed.

Constraining:	Avg Retained	#Null
Price alone	4728	0
Price + Screen + Sound	1543	13
Above + CB + PIP	951	29

When Price alone is constrained, an average of 4728 out of 10,000 draws are retained per respondent, and there is no respondent for whom no draws are retained. When all attributes but brand are constrained, only 951 of 10,000 draws are retained for the average respondent, and there are 29 respondents for whom no draws are retained. Although not reported, hit rate deteriorated with increasing numbers of constraints, probably because of the correspondingly smaller samples of acceptable draws. Performance results are in the last row of the table below

Method	Hit Rate	MAE
Unconstrained	.655	3.22
Tying Post Hoc	.659	3.22
Tying Draws Post Hoc	.664	3.31
Importance Sampling	.639	3.65

When enforcing constraints for price + screen + sound, Importance Sampling has a lower hit rate and a higher MAE than other methods. It seems likely that this is partly due to the decreased number of draws available for estimation, averaging only 1,543 per respondent instead of 10,000 available to the other methods.

Enforcing Constraints with Log-Normal Distributions

The next procedure makes use of a log-normal distributional assumption. Two changes to the CBC/HB software were required.

> First, the coding of independent variables was modified so that regression coefficients would refer to ***differences*** among levels, rather than to levels themselves. After the incremental coefficients are estimated, they are re-transformed back into conventional form.
>
> Second, at the places in the program where betas are summed to compute utilities for each alternative, the betas to be constrained are exponentiated and values of **exp(beta[i])** rather than **beta[i]** are summed. Since the transformed values are exponentials, they can never be negative. The logs of the transformed values are assumed to be normal with specified prior mean and covariances.

Results for the log-normal transformation are provided in the table below:

Method	Hit Rate	MAE
Unconstrained	.655	3.22
Tying Post Hoc	.659	3.22
Tying Draws Post Hoc	.664	3.31
Importance Sampling	.639	3.65
Log-Normal	.661	3.12

The log-normal approach produces a good hit rate and the lowest share prediction error so far. However, there is considerable uncertainty in the log-normal hit rates and MAE values. We have repeated this computation several times and noticed considerable variability from run to run. The values reported are medians for several runs. The log-normal method produces extremely large part-worth estimates for some respondents, with numerical values in the hundreds rather than with more typical single digits. Although the median results reported are impressive, the greater variability of results with this approach suggests that a less radical transformation than the exponential might be more effective.

Enforcing Constraints with "Tobit-like" Transformations

We are indebted to Kenneth Train for suggesting this approach, which has some similarities to the log-normal approach.

In "regular" CHC/HB, the "upper model" and the "lower model" use the same variables. That is, we assume each individual's part-worths are normally distributed in the upper model, and we apply a logit model to those same values in the lower model.

In the log-normal approach, by contrast, we assume the ***logs*** of constrained variables are normally distributed, and we use the anti-logs of the variables from the upper model in the logit computations for the lower model. Thus there is a change of variable between upper and lower models.

Kenneth Train has pointed out that the log transform is only one of many transformations that could be used to avoid negative values in the lower model. Another possibility is to transform each value by converting any negative value to zero. Suppose $\beta_{\mathbf{i}}$ is a normally distributed value from the upper model, and $\mathbf{b_i}$ is the transformed value to be used in the lower model, where

$$\mathbf{b_i = \max(\beta_i, 0)}$$

This "Tobit-like" transformation is less radical than the exponential transformation involved in the log-normal approach. Unlike the log-normal approach, transformed values never have larger absolute values than untransformed values. This approach produced these results:

Method	Hit Rate	MAE
Unconstrained	.655	3.22
Tying Post Hoc	.659	3.22
Tying Draws Post Hoc	.664	3.31
Importance Sampling	.639	3.65
Log-Normal	.661	3.12
Tobit-Like	.664	3.12

The last row of the table presents the best results so far, sharing highest hit rate with one method and lowest MAE with another method. However, this success comes at a considerable price: it requires re-coding the independent variables to reflect increments among underlying betas rather than using conventional effects as currently done in CBC/HB. This led to trying one more approach which we hoped might produce similar results but with conventional coding.

Trying a "Simultaneous Tying" Method

With this method we revert to effects coding as normally used in CBC/HB. However, we use a change of variable between upper and lower models similar in spirit to the Tobit-Like method. That is, we transform the variables for the purpose of computing utilities and likelihoods by tying offending values. It seemed that this approach should produce similar results, but could be implemented with less radical change to CBC/HB.

This approach is different from "Tying Draws Post Hoc." In that earlier approach, the tying is done after the estimation is concluded, and is just a way to enforce the desired order relations while producing minimal differences in the unconstrained solution. The presence of the constraints is not known to the estimation algorithm. In contrast, here the tied variables are used to compute likelihoods, so the constraints affect the solution. As with log-normal and Tobit-like solutions, we have two sets of variables: one set which is assumed to be normally distributed in the population, and a second set, obtained by transforming the first, which are used in the lower model to compute likelihoods. The upper model uses only the normally distributed variables, and the lower model uses only the transformed (tied) ones. Results for all methods tried are summarized in the table below:

TV Data Set Summary

Method	Hit Rate	MAE
Unconstrained	.655	3.22
Tying Post Hoc	.659	3.22
Tying Draws Post Hoc	.664	3.31
Importance Sampling	.639	3.65
Log-Normal	.661	3.12
Tobit-Like	.664	3.12
Simultaneous Tying	.665	3.19

"Simultaneous Tying" appears to have a slightly better hit rate and a slightly worse share prediction error than "Tobit-Like," but the differences are not significant. Standard errors were estimated for results of the last two methods by doing 50 estimations, each based on 1,000 separate HB iterations. Those standard errors are about 0.001 for hit rate and 0.05 for MAE.

Considering that it can be done with conventional coding of independent variables, these results suggest that the Simultaneous Tying method would be the best choice for implementation in CBC/HB. We next checked whether this could be verified with a second data set.

The PC Data Set

These data were contributed by Jon Pinnell from a study done in the early '90s by IntelliQuest. A sample of 326 commercial respondents considered personal computers, using 6 attributes:

Brand (5 levels)
Performance (3 levels)
How sold (3 levels)
Warranty (3 levels)
How Serviced (3 levels)
Price (5 levels)

Each respondent answered 8 customized choice tasks, consisting of 3 concepts plus a "None" option. The 5 brands chosen for each respondent were customized, consisting of his/her most and second most liked as well as several lower in preference from a longer list of brands, and arranged in that respondent's claimed order of preference.

Each respondent also answered 8 holdout tasks that were the same for all respondents, except that the respondent's customized brands were inserted in appropriate positions. Thus we are able to compute not only hit rates, but also mean absolute errors for choice share predictions. We constrained Brand, Performance, Warranty, and Price. Results for several methods are given in the table below.

PC Data Set Summary

Method	**Hit Rate**	**MAE**
Unconstrained	**.659**	**5.97**
Tying Post Hoc	**.671**	**5.68**
Tying Draws Post Hoc	**.676**	**5.42**
Importance Sampling	**.655**	**5.31**
Log-Normal	**.678**	**6.07**
Tobit-Like	**.670**	**5.60**
Simultaneous Tying	**.669**	**5.50**

Standard errors are again about 0.001 for hit rate and 0.05 for MAE, so all methods but Importance Sampling have better hit rates than the unconstrained solution, and all but log-normal have better MAEs.

The Importance Sampling method suffers from a low hit rate, probably the result of a too-small sample of draws satisfying all the constraints.

The Log-Normal method produced a high MAE, which is apparently the result of the inherent variability of its results, as seen previously with the other data set.

Curiously, the overall winner with this data set is the simple tying of draws after estimation has been done. Second place is shared by Tobit-Like and Simultaneous Tying. With this data set, Simultaneous Tying has a slightly better MAE than Tobit-Like, in contrast to the previous data set where the relationship was reversed. Our expectation that Simultaneous Tying and Tobit-Like would produce similar results appears to be validated, since they don't differ significantly in either data set, and they average out to be nearly equivalent.

Discussion

For both of these data sets, constraints almost always produced higher hit rates. They often produced smaller errors in share prediction, but there were cases where share predictions were worse. These results suggest an answer to the question of whether researchers should be concerned with monotonicity violations. If the primary purpose of the study is to predict ***individual*** choices, then it appears desirable to enforce monotonicity constraints. On the other hand, if the primary purpose of the study is to predict ***aggregate*** measures such as market shares, monotonicity constraints appear less helpful, and may occasionally even be harmful.

This conclusion was anticipated by Wittink (2000), who pointed out that constraints can be expected to reduce variance at the expense of increasing bias. He observed that hit rates are sensitive to both bias and variance, so trading a large amount of variance for a small amount of bias is likely to improve hit rates. He also observed that aggregate share predictions are mostly sensitive to bias since random error is likely to average out, so share predictions are less likely to be improved by constraints.

Of course, a potentially over-riding issue is the face-validity of results. It is important for the end-users of study results to be confident of their validity. If that confidence will be undermined by part-worths with incorrect slopes, then that alone may justify imposing monotonicity constraints.

We have decided to implement Simultaneous Tying in Version 2 of CBC/HB. However, since Tying Draws Post Hoc performs so well, we also provide a stand-alone program to produce part-worths by tying individual draws after estimation is done without constraints.

References

Allenby, Greg 2000, "Enforcing Monotonicity Constraints by Rejection Sampling," a contribution to BayesSIG, bayessig@ama.org .

Huber, Joel, Bryan Orme and Richard Miller 1999, "Dealing with Product Similarity in Conjoint Simulations," Sawtooth Software Conference Proceedings, pp. 253-266.

Orme, Bryan and Gary Baker, 2000, "Comparing Hierarchical Bayes Draws and Randomized First Choice for Conjoint Simulations," Sawtooth Software Conference Proceedings, forthcoming.

Sawtooth Software, Inc., 1999, "The CBC/HB Module," Software for Hierarchical Bayes Estimation of Conjoint Part-Worths.

Wittink, Dick R., 2000, "Predictive Validity of Conjoint Analysis," Sawtooth Software Conference Proceedings, forthcoming.