Bootstrapping of Data and Decisions

Joel Huber

The Journal of Consumer Research, Vol. 2, No. 3 (Dec., 1975), 229-234.

Stable URL:
http://links.jstor.org/sici?sici=0093-5301%28197512%292%3A3%3C229%3ABODAD%3E2.0.CO%3B2-U

The Journal of Consumer Research is currently published by Journal of Consumer Research Inc..

Bootstrapping of Data and Decisions

JOEL HUBER*

Bootstrapping is a way of improving human decisions by replacing raw judgments with a simple model of those judgments. Past work in the bootstrapping of decisions is reviewed and it is shown that the same technique can be used to upgrade the quality of data used by behavioral scientists.

Bootstrapping involves the substitution of a simple linear model of judgments in place of the judgments themselves. It has been found that in many decision making contexts the bootstrapped decisions are better than the judgments from which they were derived. It appears that the linear model is quite successful at capturing the policy of the judge and then making decisions without human inconsistency. Most of the work done on bootstrapping has been done in a context—such as forecasts—where the criterion or accuracy is clearly defined. This paper reviews past work done in bootstrapping and shows that it can be used to upgrade the quality of subjective judgments (data). These judgments have no ultimate criterion of accuracy but are evaluated in terms of their usefulness as input to a predictive model. Implications are explored as to the use of bootstrapping of both data and decisions in consumer behavior.

Bootstrapping of decisions has generally taken the following paradigm. Subjects are given sets of cues by which to make decisions. These might be test scores and evaluations used to determine college admissions. A simple additive or linear model is formed predicting judgments as a function of the cues. When a suitable criterion for accuracy exists, such as eventual rank in class, the simple model has generally been found to be more accurate than the raw judgments. That is, the correlation of the ultimate criterion with the model is generally greater than its correlation with the raw judgments used to derive the model. The replacement of raw judgments with this linear combination of cues has come to be called "bootstrapping," a term coined by R. M. Dawes. In effect, the raw judgments "lift themselves up by their bootstraps." Two examples illustrate this technique.

One of the first studies to suggest that simple models of decisions might produce good decisions was Yntema and Torgerson (1961). Subjects were provided with ellipses of different size, shape and color. Worth was defined so as to increase nonlinearly with increases in size, thinness and brownness. After a ten day training period consisting of giving subjects feedback on their predictions of worth, subjects were required to make a battery of 180 judgments without feedback. The average product moment correlation between these judgments and the true worths was 0.84. A simple additive bootstrapping model derived from predicting these judgments as functions of size, shape and color yielded an average correlation with the true values of 0.89. Thus, the additive bootstrapping models were more accurate than the judges in spite of the fact that these models could not take into account interactions, while the human judges presumably could.

Goldberg (1970) used judgments of clinical psychologists to build a model to discriminate neurotics from psychotics. The predictor variables were scores on the Minnesota Multiphasic Personality Inventory, a test which provides a profile of patients along 11 dimensions. The bootstrapping model provided superior predictions of later diagnosis for 26 out of the 29 judges. Similarly, Dawes (1971) found that admissions evaluations have a higher correlation with actual achievement if derived from a bootstrapping model of judgments rather than the judgments themselves. Bowman (1963) and Kunreuther (1969) were able to demonstrate improved decisions in the field of production management; while in marketing, Heeler *et al.* (1973) and Montgomery (1972) have modeled the decisions of buyers for supermarket chains with similar results.

WHY BOOTSTRAPPING WORKS

Bootstrapping works because the linear model is able to make extremely good approximations of most decision processes. The model then makes these judgments without random error. Thus by bootstrapping one replaces the random error of the judge with the nonrandom error of the model. This nonrandom error can be broken into two components: (1) a calibration

* Joel Huber is Assistant Professor of Industrial Administration, Krannert School of Industrial Administration, Purdue University.

error due to insufficient sample size to estimate reliably the parameters or (2) a specification error due to the inability of the linear model to capture the complexities of what the judge is doing. These two sources of error will be discussed in an attempt to explain why their sum has generally been less than the error of raw judgments.

Calibration error reflects a fairly minor component of the error in a bootstrapping model. Dudycha and Naylor (1966) provided subjects with cues and worths that were related by a linear model with different levels of error. After a learning period, the bootstrapped models of judgments on 50 stimuli had average correlations with the optimal model of better than 0.90. Even when the beta coefficients appear to be unstable themselves due to multicollinearity, the predictions from such a model tend to be quite stable.

Specification error reflects the inability of the linear model to account for nonlinearities or interactions in the judgment process. The early researchers of paramorphic representation (concentrated mainly at Oregon Research Institute) saw the linear model as a first approximation to the decision process which would be modified later by nonlinear and interactive components. Then a funny thing happened. Adjustments to the linear model provided very little improvement to predictive accuracy. This result was anticipated by Yntema and Torgerson (1961) who found that a main effects model $Y = a_i + b_j + c_k$ accounts for over 90 percent of the variance of data generated by the multiplicative model $Y = ij + ik + jk$ (where i, j, and k are integers between one and seven). In general, one loses very little by approximating such a decision process with only the main effects.

In a large simulation study Rorer (1971) tested the ability of the linear model to approximate data generated by interactive and configural models. These included interactive and configural terms as well as disjunctive and conjunctive step functions and elaborate lexicographic models. The linear model was generally able to account for over 70 percent of the variance. Furthermore, given the level of error typically found in human judgment, the interaction term would not be significant. With such data the analysis of variance generally lacks the power to measure interactions, even where they, in fact, exist. This result appears to be quite general as long as *the criterion variable is conditionally monotone with respect to the cues.* That is, if the direction of the effect of a cue is the same regardless of the levels of the other cues. Cues which are conditionally monotone appear quite often in judgmental situations. For example, economy, performance, styling and closeness to mid-sized are all attributes which in a rational man might be conditionally monotone with his judgments of overall worth of automobiles. Research has shown that if this is the case then a linear model will do a good job of approximating the judgments.

While the bootstrapping of decisions generally produces better results than the decisions from which it was derived, it generally does not produce the best linear decision scheme available. Simply regressing the cues directly on the criterion produces better results than going through judgments (Meehl, 1954). In fact, Dawes and Corrigan (1974) show that linear models with random coefficients (but the correct sign) do as well as the bootstrapping models. Thus bootstrapping models are not to be seen as magical or in any sense optimal linear models, but merely a method for picking the appropriate variables and weighing them in the right direction. Furthermore, if an unambiguous criterion exists, a better model can be derived by regressing the cues directly on it.

It could be argued that there are many situations where the "optimal" linear model derived from a regression of cues on the criterion is less valid than the bootstrapping model derived from a regression of cues on decisions. Consider the admissions problem. Bootstrapping does not do as well as an optimal linear model in predicting rank in class. However, it is feasible that the admissions committee is taking into consideration other goals, such as racial balance or being well-rounded. These considerations would be reflected in the coefficients of the bootstrapping model but not in a model calibrated to class rank. Thus if the objective is to provide a model that satisfies the judge's implicit goals, bootstrapping provides at least a first step in this direction. It is this quality that makes bootstrapping particularly appropriate to many problems in consumer behavior.

DATA BOOTSTRAPPING

The typical validations of bootstrapping have used judgments, such as forecasts, for which the ultimate criterion for accuracy is easily specified. Further, the cues that go into the judgments have been clearly specified and known to the subject. This study considers the applicability of bootstrapping to data that serve as input to behavioral models and generally lack the above qualities. A response of a subject to a stimulus or question cannot have ultimate validity but only be considered better or worse to the extent that it can be related to other responses or behavior on the part of the subject. For example, the superiority of a measure of intention to purchase can be ascertained on the basis of its correlation with actual purchases. In the same way bootstrapping will be evaluated on the basis of its effectiveness in improving input to a behavioral model.

The present study represents an attempt to evaluate bootstrapping on preference judgments of particular

samples of iced tea. All analysis is done on the basis of the individual subject. The test between bootstrapped and raw judgments is made by comparing which provides better predictions of preference.

Input and Predictive Models

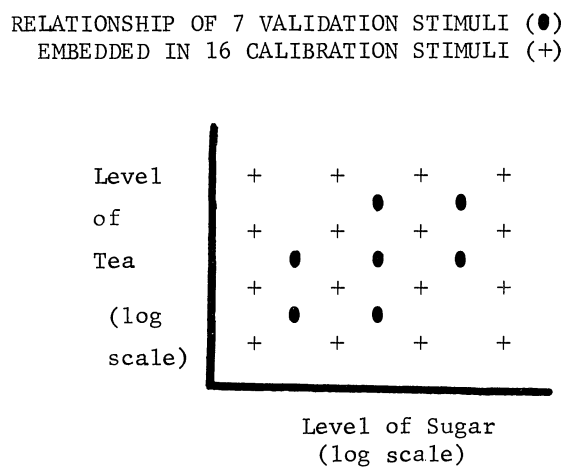
The preference judgments from a convenience sample of 22 people were used for this study. Each was required to make judgments on samples of Lipton iced tea that differed in the amount of sugar and tea according to a balanced design. As is illustrated in Figure 1, they were required to make independent judgments on 7 validation stimuli and 16 calibration stimuli.

For each subject the analysis revolved about the following data.

P_i = Preference scale for stimulus i, i-1,16 for the calibration stimuli and i-17,23 for the validation stimuli. This scale was formed for each set from preference differences using Scheffé's (1951) method of analysis modified for analysis of individual data.

δ_{ik} = Judgment as to the degree to which stimulus i has too much, or too little, sugar ($k = 1$) or tea ($k = 2$). These were coded on an integer scale from -3 to $+3$, negative numbers indicating too little, zero indicating optimum, and positive numbers indicating too much of the ingredient.

x_{ik} = Objective level of sugar ($k = 1$) and tea ($k = 2$) for stimuli i.



Preference for iced tea of 16 stimuli (+) are used to calibrate models to predict individual preference on 7 valication stimuli (●).

FIGURE 1

Using data on the calibration stimuli, preference for each subject

$$P_i = f(\delta_{ik}) = b_o + b_1 |\delta_{i1}| + b_2|\delta_{i2}| \qquad (1)$$

is estimated by multiple regression. This is a version of the familiar weighted-additive model. The absolute value of the δ_{ik}'s can be interpreted as the distance from stimulus i from the ideal along dimension k. Thus preference is assumed to be a function of the sum of weighted distances along these psychological dimensions.

The bootstrapping model relates the δ_{ik}'s to the real levels of sugar and tea. This is for dimension k

$$\delta_{ik} = g(x_{ik}) = b_{ok} + b_{1k}x_{i1} + b_{2k}x_{i2}. \qquad (2)$$

The bootstrapping model given in this equation assumes the amount of change desired in tea and sugar in a linear function of the actual values of these variables. If, for a given individual, preferences are single-peaked or monotone within the physical space, then the physical levels will be monotone with the δ_{ik}'s. Furthermore, since linear functions have been shown (Rorer, 1971, Dawes and Corrigan, 1974) to produce close approximation to most monotone functions, the linear model appears reasonable in this case. This is further supported by the fit of the calibration stimuli to Equation 2. The average product moment correlation across subjects was .84 for sweetness and .62 for tea.

A Test of Data Bootstrapping

The data to be bootstrapped are the subjective estimates of sugar and tea (δ_{ik}'s). These are related to actual levels of sugar and tea in Equation 2 ($\delta_{ik} = g(x_{ik})$). Bootstrapping is tested by comparing the effectiveness of the predicted data against the raw data as input to the preference equation ($P_i = f(\delta_{ik})$). This comparison is made at two junctures of the prediction process: (1) to parameterize the preference equation and (2) as input to the parameterized model. In both cases the predicted δ_{ik}'s are simply substituted for the raw δ_{ik}'s to test bootstrapping. The criterion is which input provides better predictions of the validation stimuli.

As is shown in Figure 2, using bootstrapping to parameterize has relatively little effect. However, its use on models that have been parameterized produces large and significant gains in prediction.

If one considers the bootstrapping equations to be the first stage in a two-stage model, then using bootstrapping to parameterize the preference model is equivalent to two-stage least squares. This procedure has some theoretical advantages in that errors of the bootstrapped values are not correlated with the error terms of the preference scores. In this case, however, the two-stage model did not produce significant gains

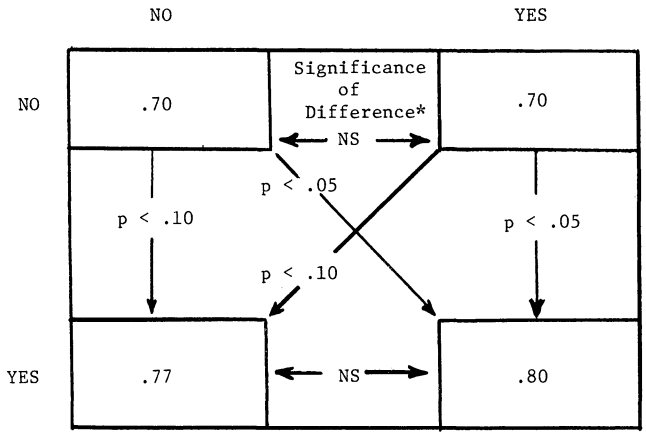


FIGURE 2
Effect on Prediction of Using Bootstrapping Measures of Subjective Sugar and Tea for Parameterizing and as Input to Predictive Model

Root-mean-square of product moment correlation of predicted to actual preference scale.

Bootstrapping Used to Parameterize

Bootstrapping Used in Model for Prediction

* The significance of the difference in the predictive fit of a pair of models is done by taking the difference between the Fisher (1922) z-transform of the correlations for each subject. A t-test is then used to test the null hypothesis that this difference is uniformly zero across the 22 subjects in the study.

probably because the errors in the δ_{ik}'s are relatively random and because of the well-known robustness of linear regression to random error in the independent variables.

By contrast, using bootstrapping to produce variables as input to the parameterized models produced large gains in predicting the preferences on the validation stimuli. This result could only have occurred if the bootstrapped values were, in fact, more accurate estimates of subjective sugar and tea than the original data.

DISCUSSION

Bootstrapping could be used to upgrade the quality of data in a variety of behavioral science contexts. For example, judgments of sportiness in automobiles could be modeled as linear functions of speed, acceleration, width-to-height ratio, and cornering ability. As input to a second model involving product choice these bootstrapping models would provide not only greater reliability but also assistance to designers interested in translating "sportiness" into product specifications.

In another context subjects could be asked to provide preferences on houses built to various specifications. If the specifications form a balanced design over the different houses, then it is possible to derive an additive model that decomposes total preference into a utility value for various levels of each attribute. It is hypothesized that the predicted preferences would be a better predictor of later choice between stimuli than the raw preferences.

The use of a model to impart reliability to raw data is hardly novel in behavioral research. The technique perhaps most closely related to the above use of data bootstrapping is spatial or temporal smoothing. Instead of assuming a global linear relationship, these techniques assume local linearity so that each point can be approximated by a simple function of several contingent points. MacKay (1973) used spatial smoothing to impart greater reliability to store-usage data and found that this resulted in better fits and more interpretable solutions to the quadratic regressions which were used to produce market penetration maps. Even in cases where there is not an objective physical or temporal dimensionality, the assumption of linearity is used to improve data quality. Factor analysis assumes a linear relationship between variables and uses this to impart redundancy in the factors. In the same way the bootstrapping equations used in this study employ a particular linearity assumption to increase the reliability of any particular judgment by forcing it to be consistent with the other judgments in the set.

There is a simple preliminary test to determine whether a bootstrapping model will upgrade the quality of data. It requires one replication of the original judgments. A bootstrapping model calibrated on each half is validated against the other half. The average correlation is called the "double cross-validated correlation." This is compared to the reliability of the judgments which is simply the correlation between the two halves. If the model is exact, the cross-validated correlation should approach the square-root of the reliability. It will not be exactly so unless the sample of judgments is infinite. For practical purposes, however, if the cross-validated correlation is greater than the reliability of the judgments, then the bootstrapping model is better at predicting the raw data itself. There are some risks attached to this procedure. One is that the two replications may not be independent draws from the same judgment process but may reflect a change of viewpoint or merely a repetition of the first judgments. The second problem is that the bootstrapped relationship, however optimal it might be with respect to the judgments, may not be at all optimal with respect to the criterion or use to which it is being put. In Einhorn's (1970) study, the optimal combination of cues to predict the criterion was disjunctive while the optimal bootstrapping model was conjunctive.

To avoid just these kinds of problems Kunreuther (1969) advocates the use of bootstrapping only when (1) the decision rules are constant over time, (2) the model and resultant coefficients make theoretical sense and, (3) are statistically significant. While these rules

apply more in a production programming context than the present one, certainly the call for *a priori* consideration of the bootstrapping model is valid.

IMPLICATIONS FOR CONSUMER BEHAVIOR

The decompositional models that have been considered here as bootstrapping models are likely to be as good at "capturing" consumer decisions as they have been of capturing the decisions of admissions committees, psychologists and production planners. It is also likely that the same frustrating search for interactions and mediator variables that has occurred in these areas will be paralleled in research on consumer behavior.

Some of these potential difficulties in uncovering interactions and nonlinearities can be illustrated in the problem of discriminating between judgments produced by conjunctive and disjunctive decision rules. Consider the conjunctive decision rule. Here a judge finds a stimulus more acceptable if it is high with respect to all attributes. The isopreference (or isoacceptance) contours are convex to the origin. By contrast, the disjunctive decision rule accepts a stimulus if it is high on at least one attribute, and thus produces concave isopreference contours. These are illustrated in Figure 3.

The problem of discriminating between the two rules occurs when the cues are themselves positively correlated. This commonly happens in real situations and means that most of the data points are in Regions 2 and 4. Unfortunately these regions result in identical decisions under both rules. The answer to this problem is, as advocated by Einhorn (1970) and Anderson (1970), to provide cues in a balanced design so that Regions 1 and 3 have enough data points to discriminate between the models. There are two problems associated with this strategy, however. Since, due to their positive correlation, cues in these critical regions are rare, decisions on such unfamiliar combinations may be difficult. Furthermore, the conjunction of certain cues might be considered simply unbelievable. In either case the judge's strategy might change and thus not reflect decision rules in other regions of the space. The second problem occurs whenever a judge is required to do a series of judgments according to a pattern of cues. At least one conceivable response is for the subject to determine a simple decision rule that enables quick completion of the experimental task. This decision rule might or might not reflect the judge's decisions in the field. Thus the researcher of decision processes is in a familiar dilemma. The researcher must introduce artificial or unfamiliar stimuli in order to discriminate between models, but this artificiality leads to uncertainty as to whether the results reflect authentic decision behavior or merely taskmanship on the part of the subject. It is likely that future researchers in consumer behavior will not be able to get between the horns of this dilemma but will continue to be impaled on one tip or the other.

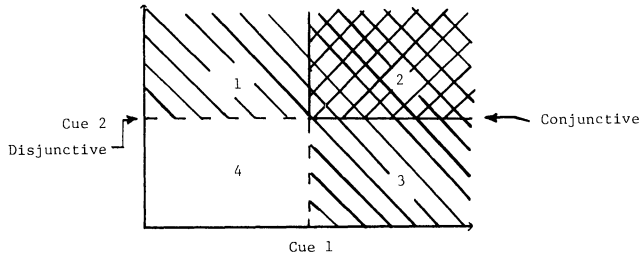


FIGURE 3
Isoacceptance Curves Under Different Decision Rules

Region 2 accepts under conjunctive rule, while region 1, 2 or 3 accept under the disjunctive rule.

CONCLUSION

The user of bootstrapping, to the extent that he is more concerned with approximating than understanding decision processes, is content to use the robust linear models and rest secure in the knowledge that rare combinations of cues and nonlinearities have little effect on explained variance. Such predictive, decompositional models can, however, be quite useful in the study of consumer behavior. They could be used as a preliminary normative step to enable the consumer to understand the implications of his own decision processes. Alternatively, they could be used by regulatory agencies as a first approximation of what the consumer decision process is. Values then might be inferred from decisions rather than imposed from above.

In contrast to bootstrapping of decisions, bootstrapping of data is likely to have more limited use in consumer behavior. It can be seen as one of a number of methods to provide reliability in data through the structure of a model. However, just as such models should be consistent with existing theory, they can add to it by explaining, at least in a preliminary way, the sources of the raw judgments. Thus, data bootstrapping can provide more than just reliable data, it can provide a basis for an understanding of its own validity.

REFERENCES

Anderson, Norman. "Functional Measurement and Psychophysical Judgment," *Psychological Review,* 77 (1970), 153-170.

Bowman, E. H. "Consistency and Optimality in Management Decision Making," *Management Science,* 9 (1963), 310-321.

Dawes, Robyn M. "A Case Study of Graduate Admissions: Application of Three Principles of Human Decision Making," *American Psychologist,* 26 (1971), 180-188.

Dawes, Robyn M. and Bernard Corrigan. "Linear Models in Decision Making," *Psychological Bulletin,* 18 (1974), 95-106.

Dudycha, L. W. and J. C. Naylor. "Characteristics of the Human Inference Process in Complex Behavior Situations," *Organizational Behavior and Human Performance,* 1 (1966), 110-128.

Einhorn, Hillel J. "The Use of Nonlinear, Noncompensatory Models in Decision Making," *Psychological Bulletin,* 73 (1970), 221-230.

Fisher, R. A. "The Goodness of Fit and Regression Formulae, and the Distribution of Regression Coefficients," *Journal of the Royal Statistical Society,* 85 (1922).

Goldberg, L. R. "Man Versus Model of Man: A Rationale, Plus Some Evidence, for a Method of Improving on Clinical Inferences," *Psychological Bulletin,* 73 (1970), 422-432.

Heeler, Roger M., Michael J. Kearney and Bruce McHaffey. "Modeling Supermarket Product Selection," *Journal of Marketing Research,* 10 (February 1973), 34-37.

Hoffman, P. J. "The Paramorphic Representation of Clinical Judgment," *Psychological Bulletin,* 57 (1960), 116-131.

Kunreuther, Howard. "Extensions of Bowman's Theory of Managerial Decision Making," *Management Science,* 15 (April 1969), 415-439.

MacKay, David B. "Spatial Measurement of Retail Store Demand," *Journal of Marketing Research,* 10 (November 1973), 447-453.

Meehl, P. E. *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of the Literature.* Minneapolis: University of Minnesota Press, 1954.

Montgomery, David B. "New Product Distribution: An Analysis of Supermarket Buyer Decisions," Marketing Science Institute Research Programs: Cambridge, Mass., 63 (1973).

Rorer, L. G. "A Circuitous Route To Bootstrapping." In H. B. Haley, A. G. D'Costa and A. M. Schafer (Eds.) *Conference on Personality Measurement in Medical Education,* Wash., D.C., Associations of American Medical Colleges, 1971.

Scheffé, H. "An Analysis of Variance For Paired Comparisons," *Journal of the American Statistical Association,* 47 (1952), 381-400.

Yntema, D. B. and W. S. Torgerson. "Man-Computer Cooperation in Decisions Requiring Common Sense," *IRE Transactions of the Professional Group on Human Factors in Electronics,* HFE-2(1) (1961), 20-26.