APPLICATIONS OF THE OVERLAPPING INFORMATION MODEL

Stanley L. Warner, York University

1. INTRODUCTION

A problem of importance in every organization is how to secure good summary information for particular decisions. A summary should be balanced and comprehensive, but such characteristics are difficult to measure. Consequently it is difficult for an organization to tell whether or not summaries produced by its staff are adequate. Since even a slightly better summary at a critical time may prevent a costly error, even slight improvements in the ability to measure and monitor summary information may be of value.

The overlapping information model provides an approach to measuring characteristics of summary information through measuring the influence of the summary on persons chosen as a test population. Supposing the summary is designed to present information for and against some proposition, the model suggests interpretations for simple statistics based on recording the personal opinions of the test population both before and after presentation of the summary. While some measurements will depend on the choice of the test population, some useful inferences regarding the summary appear to be possible provided that the members chosen for the test population process information in at least a roughly coherent Bayesian fashion. In particular, the assumption that those in the test population at least partially discount new information which overlaps information that they have seen before suggests the possibility of measuring certain characteristics of the summary, the summarizers, and members of the test population.

2. THE INFORMATION MODEL

The purpose of the summary is to condense a given body of facts and arguments concerning whether or not some hypothesis is true. For the model, in analogy with finite sampling procedures, the collection of relevant facts and arguments to be summarized is considered as if it were a population of data, each piece of which can influence personal subjective probabilities of belief in the hypothesis. The summarizer is charged with selecting an influential subset from the population of data in as representative a fashion as possible, the objective being to approximate both the total positive and the total negative influence that would be conveyed if the entire population of data could be presented and assimilated. The resulting summary is presented independently to the test population members, who, in analogy with the Bayesian processing of information, are expected to modify prior beliefs expressed as probabilities to posterior beliefs expressed as probilities.

Under ordinary circumstances the members of the test population would not be the final decision-makers, but merely a group of persons thought to process information rationally so that their reported prior and posterior opinions could be used to help evaluate the summary information. The model is thus concerned with the final decision only indirectly, in that it is concerned with the quality of the summarized information which is provided to the decision-makers. Neither the final decision-makers nor the final decision are a part of the model.

Since randomized selection represents a widely accepted standard for judging impartiality in selection, the model considers implications suggested by randomized selection of summary data from the hypothetical population of data. In an application, the objective would be to have the summary, even though necessarily purposefully chosen, sufficiently representative of the population to satisfy measurable criteria for balance suggested by assuming randomized selection.

To represent the procedure, the opinion of the ith member of the population prior to receiving the summarized data is represented by P_i, and after receiving the data by Q_i, $i = 1,2,\ldots,n$. The net influence of the summary on the ith person is measured by

$$I_i = \ln(Q_i/(1-Q_i)) - \ln(P_i/(1-P_i)). \qquad (1)$$

All relevant facts and arguments are viewed as if conceptually combined into a population of N pieces of data, each of the N pieces being so constituted as to be a relatively independent piece of influence regarding the proposition. The influence of data in information terms is thus additive, so that, regarding Y_{ij} as the information seen by the ith person in the jth piece of data, the prior ln odds of person i is represented as

$$Z_i = \ln(P_i/(1-P_i)) = \sum_{j\ \varepsilon\ A(i)} Y_{ij} \qquad (2)$$

where $j \varepsilon A(i)$ refers to the k_i pieces of data that have been seen by person i before the summary.

Similarly, with S_i representing the sum of the Y_{ij} over the m units presented in the summary, and V_i representing the sum of the Y_{ij} over the x_i overlapping units that are both in the summary and the set seen before by person i, the basic identity for the model is given by

$$I_i = S_i + D_i V_i \qquad (3)$$

with $D_i \geq -1$. Defining $\bar{Z}_i = Z_i/k_i$, $\bar{V}_i = V_i/x_i$, $r_i = x_i/k_i$, and $U_i = k_i(\bar{Z}_i - \bar{V}_i)$, the basic identity of (3) can also be written as

$$I_i = S_i + D_i r_i (Z_i - U_i) \qquad (4)$$

where U_i is considered a measurement error with expected value 0 relating the unobservable $k_i\bar{V}_i$ to the observable Z_i. An earlier result in Warner (1984) considered inferences from the simple regression model implied by assuming all $D_i = -1$, noting randomized selection implies all $Er_i = m/N$, and ignoring all bias for simplicity.

Somewhat more realistic assumptions that still allow useful inferences are given by allowing Z_i to be positively correlated with U_i, S_i to be positively correlated with Z_i but not correlated with U_i or $D_i r_i$, and considering $D_i r_i$ independent

of Z_i and U_i with D_i and r_i uncorrelated.

In particular, defining LSB as $COV(I,Z)/VAR(Z) - EDEr$, $\beta_i = D_i r_i + LSB$, and $\alpha_i = S_i - D_i r_i U_i - LSB\ Z_i$, the next section shows that the random coefficients model

$$I_i = \alpha_i + \beta_i Z_i \tag{5}$$

may be estimated by conventional methods. In particular, familiar estimates of the expected values and variances of the α_i and β_i are given by Hildreth and Houck (1968), and for the individual coefficients α_i and β_i by Griffiths (1972). For these estimates the Z_i are to be taken as given, and a more complete model would require the specification of how the Z_i are themselves determined. It is also emphasized that, while an attempt has been made to allow for the more obvious correlations among the elements of the model, other possible correlations make practical inference hazardous. The next section demonstrates at least the possibility of inferences through relating the listed assumptions of the information model with the usual assumptions of the random coefficients estimation model.

3. THE RANDOM COEFFICIENTS ESTIMATION MODEL

Toward interpreting estimates of (5) it is first to be noted that with

$$LSB = COV(I,Z)/VAR(Z) - EDEr, \tag{6}$$

$$\beta_i = D_i r_i + LSB, \tag{7}$$

and

$$\alpha_i = S_i - D_i r_i U_i - LSB\ Z_i, \tag{8}$$

the model parameters

$$E\alpha_i = ES - LSB\ EZ, \tag{9}$$

$$E\beta_i = EDEr + LSB, \tag{10}$$

and

$$COV(\alpha_i \beta_i) = E(S_i - D_i r_i U_i - LSB(Z_i))(D_i r_i - EDEr) = 0 \tag{11}$$

under the assumptions.

In terms of the explanatory variables, approximately in large samples, the identity of (4) shows

$$COV(I,Z)/VAR\ Z = E(S_i + D_i r_i Z_i - D_i r_i U_i)(Z_i - EZ)/VAR\ Z$$

$$= COV(S,Z)/VAR\ Z + EDEr - EDEr\ COV(U,Z)/VAR\ Z$$

so the least squares bias is

$$LSB = COV(S,Z)/VAR\ Z - EDEr\ COV(U,Z)/VAR\ Z. \tag{12}$$

For applications it is thus to be noted that, if $ED < 0$, LSB will be > 0 since the other values appearing in (12) are all > 0 by assumption.

That the usual estimates of the parameters $E\alpha_i$ and $E\beta_i$ will be consistent under the assumptions can be seen by writing

$$I_i = E\alpha_i + (E\beta_i)\ Z_i + w_i \tag{13}$$

with $w_i = \alpha_i - E\alpha_i + (\beta_i - E\beta_i)\ Z_i$.

Thus, approximately in large samples,

$$Ew_i(Z_i - EZ) = E(S_i - D_i r_i U_i - LSB\ Z_i)(Z_i - EZ)$$

$$+ E(D_i r_i - EDEr)(Z_i)(Z_i - EZ)$$

$$= COV(S,Z) - EDEr\ COV(U,Z) - LSB\ VAR\ Z + 0$$

with the last zero the result of the independence assumptions between $D_i r_i$ and Z_i. The expression for LSB in (12) thus shows the remaining terms reduce to zero in large samples so

$$Ew_i\ (Z_i - EZ) = 0 \tag{14}$$

and generalized least squares estimates are consistent.

4. APPLICATIONS AND INFERENCES

Supposing the assumptions of the model are reasonable approximations, estimates of the parameters based on the last section imply bounds on several concepts of interest regarding the summary. For example, since Er is bounded by 0 and 1 and LSB is positive, the sign and a range of values for ED is suggested by the estimate of $E\beta_i$. Each specifically assumed pair of values for ED and LSB provide information regarding the completeness of the summary, measured by Er; the effect of the summary on a hypothetical person with no prior information, measured by the vertical intercept; and the effect the population of data would have if it were seen, measured by the horizontal intercept. Since the Z_i are assumed unrelated to the D_i, information regarding the balance of the summary is suggested by comparing implications for Er based on computing separate regressions using observations with small Z_i and observations with large values of Z_i. Comparisons between the D_i for different members of the test population are possible through comparing estimated values for the β_i, and comparisons between different summaries or summarizers are possible through comparing parameters estimated with different summaries and randomly drawn subsets from the test population.

It is to be emphasized that all inferences from this or similar information models must be carefully qualified. Some of the information parameters being estimated, such as the relative balance of different summarizers, may be estimated through more direct experimental design that does not place so much weight on the assumptions of the model. Some balanced information experiments of this type are described in Warner (1975, 1981). The advantage of the simple model of this paper is that it is remarkably easy to apply. The next section reports an example.

5. AN EMPIRICAL ILLUSTRATION

To demonstrate the method, in February of 1985 a telephone survey of Carleton University students was arranged through the Carleton Journalism Poll

in Ottawa, Canada. The interviewees were first asked for their opinion regarding whether of not they thought that an elected Canadian Senate would be preferable to the existing appointed Senate. They were next asked if they could express their degree of belief in percentage terms, analagous to the percentage terms commonly used by weather forecasters to express their degree of belief in propositions such as "it will rain tomorrow." The reported numbers ranged from 0 to 100 and after being divided by 100 were interpreted as the probabilities P_i required by the model.

The interviewees were then asked if they would be willing to hear a short summary of a television debate that had previously taken place. A brief six sentence summary was then presented over the phone, and all interviewees were asked if they would care to modify their original percentage reply in light of the summary. The resulting numbers were converted to numbers between 0 and 1 as before and interpreted as the probabilities Q_i required by the model.

A total of 417 students participated in the study, and 316 replies were realized in which neither the first nor the second probabilities took the values of 0 or 1. For the purpose of the model, the numbers 0 and 1 are illogical because they imply that no additional information could change the degree of belief. Of the 316, there were 163, or slightly over half, who changed their reported probabilities of belief after hearing the summary. The average ln odds after the summary was virtually the same as that before the summary, with the variance slightly smaller.

Turning to the regressions, the Hildreth and Houck generalized least squares estimates for $E\alpha_i$ and $E\beta_i$ were 0.125 and -0.253 with estimated standard errors of 0.038 and 0.040; the estimates of $VAR\alpha_i$ and $VAR\beta_i$ were 0.256 and 0.081. The important quantity represented by the β_i thus was estimated to have a negative expectation and to have a relatively small variance. This is consistent with the assumption that those in the test population did at least partly discount information they had seen before. The estimates of the individual coefficients computed according to Griffiths (1972) were virtually all between -1 and 0.

Under the assumptions of the model the discount factor D_i is not related to the previously seen information indexed by the Z_i. Thus, differences in estimates of the slope parameter for small Z_i and for large Z_i may provide some evidence of the balance in the summary, since the slope is a product of the discount factor and the overlapping information proportion. It is thus of some interest that a regression using observations defined by the lowest half of the Z_i resulted in an estimate for the slope of -0.31 with a standard error of 0.08, while a regression using observations defined by the highest half of the Z_i resulted in an estimate for the slope of -0.16 with a standard error of 0.10. This suggests the possibility that in the summary the positive information was less well represented than the negative information.

REFERENCES

Griffiths, W. E., (1972), "Estimating Actual Response Coefficients in the Hildreth-Houck Random Coefficient Model," *Journal of the American Statistical Association*, 67, 628-632.

Hildreth, Clifford, and Houck, James P., (1968), "Some Estimators for a Linear Model With Random Coefficients," *Journal of the American Statistical Association*, 63, 584-595.

Warner, Stanley L., (1975), "Advocate Scoring for Unbiased Information," *Journal of the American Statistical Association*, 70, 15-22.

Warner, Stanley L., (1981), "Balanced Information: The Pickering Airport Experiment," *The Review of Economics and Statistics*, 63, 256-262.

Warner, Stanley L., (1984), "The Overlapping Information Model for Evaluating Summary Information," *Proceedings of the Social Statistics Section of the American Statistical Association meetings*, August 1984, Philadelphia, Pennsylvania, 581-584.