AN IMPUTATION PROCEDURE FOR DETERMINING MISSING FACTOR LEVELS IN ANALYSIS OF VARIANCE

Ronny A. Schaul, Bureau of the Census
Jack C. Hayya, The Pennsylvania State University

A problem of missing independent variables arises in the processing of questionnaire data. Such a problem has occurred in the construction of a price index for new housing. The price of the house is the dependent variable, Y. *The independent variables,* X_i *are characteristics (square feet of floor, number of bathrooms, air-conditioning, etc.) consisting of several levels. The Construction Statistics Division of the Bureau of the Census obtains this information from a questionnaire survey. In some questionnaires only partial information is returned. A solution is to "guess" the missing levels of the* X_i *by classifying the dependent variable with the incomplete information on the* X_i *into one of several populations. The net effect in regression, for example, is to reduce the variance of the regression coefficients, thereby reducing the variance of the predicted* $\hat{Y}$.

I. INTRODUCTION

In a questionnaire survey, it may happen that some respondents do not complete all the items on the questionnaire. Thus some questionnaires with missing information are returned; and even though a follow-up may be initiated, there is no guarantee of obtaining the missing data when the respondent considers them confidential or proprietary (or he may simply be uninterested). But, in order for these questionnaires to be useful, all items in them must be completed. This happens, for example, when the questionnaire survey is used for regression purposes.

Given that certain necessary information is already available on an incomplete questionnaire, it seems desirable to impute the missing answers in order to reduce the loss of information. The imputation procedure considered in this paper requires that the items imputed can assume only a finite number of values; i.e., the imputation procedure is applicable in an ANOVA situation.

We give two examples from surveys conducted by the Construction Statistics Division of the Census Bureau. One example deals with a survey to determine a price index of new one family houses sold; the other deals with a survey to determine a cost index of residential buildings with two or more housing units.

A Price Index of New One Family Houses Sold

In order to establish a price index, respondents to a survey are asked to supply data concerning the following nine items:

1. The price of the house;
2. The size of the house (in terms of square feet of floor area grouped into 9 classes);
3. The number of stories;
4. The number of bathrooms;
5. The presence or absence of air-conditioning;
6. The type of parking facility (garage or no garage);
7. The type of foundation (basement or no basement);
8. The geographic location (in terms of 12 areas);
9. The metropolitan location (inside or outside the Standard Metropolitan Statistical Areas as defined by the Office of Management and Budget).

Items 2 to 9 above refer to eight characteristics of the house. Information regarding items 8 and 9 is never missing, since this information can be determined independently of the respondent. Consequently, the interest is in imputing a maximum of six characteristics.

In practice, about six percent of all questionnaires have at least one of the items 2 to 7 unanswered. Since the price of the house is practically a continuous variable, imputations for missing prices will not be considered here. Also since the price of the house is regressed against all eight characteristics, an imputation procedure for missing values of these characteristics seems desirable, especially when a questionnaire has only one or two entries missing.

In general, the nonresponse conforms to the following distribution:

Square feet only	32.0%
Basement only	6.1
Air-conditioning (AC) only	26.4
Garage only	6.1
Stories only	2.0
Bathrooms only	1.5
AC and garage	0.5
Garage and basement	0.5
AC, garage, and basement	1.5
All six characteristics missing	10.2
All other combinations	13.2
	100.0%

A Cost Index of Residential Buildings With Two or More Housing Units

Questionnaires similar to those for deter-

mining a price index of new one family houses are used here. About 40% of all questionnaires have at least one of the requested items on the characteristics of the buildings unanswered.

Review of Previous Work

Imputation procedures for similar problems have been used by the Bureau of the Census (See Chapman, [2, pp. 27-]). As described by Chapman, two main procedures are the Cold Deck procedure and the Hot Deck procedure. Both classify the data into cells according to the characteristics so that "responses will be relatively homogeneous within cells and heterogeneous between cells... For each missing item for a particular respondent to the...survey, the values of the appropriate completed items are noted to identify the relevant cell. The respondent is associated with the cell corresponding to the values of the items. A value is then selected from the responses in the cold deck included in the same cell. This value is usually selected at random or systematically" (Chapman, [2, p. 4]). Variations of this method consist in using "a moving average of values in a cell to substitute for a missing value," or "an imputed value...obtained from a regression of the particular item on several of the other items." (Chapman, [2, p. 6]). The difference between the two methods is that the Cold Deck procedure uses data from a previous survey, whereas the Hot Deck procedure uses data from the same survey. The interested reader may also refer to [4, 6, 7, 8, 9].

Currently, no imputation procedure is used by the Bureau of the Census in either of the two examples mentioned above. Incomplete questionnaires are not used when calculating the indices. It is planned to use the procedure proposed in this paper if the imputation improves the estimates of the regression coefficients which are used when computing the indices. Whether improvement is achieved or not is to be determined by simulation and by comparing the regression using the complete data set on the one hand, and the regression using the complete data set augmented by imputation on the other.

II. THE IMPUTATION PROCEDURE

The following is a modification of the Hot Deck procedure. It takes advantage of the fact that the variables for which imputations are made can assume only a finite number of values.

Determining Factors and Factor Levels

Suppose we have a factorial design with I factors, where factor K has L_K levels, $1 \leq K \leq I$. For instance, in the first example given in section I (the price index for new one family houses sold), the number of factors is I = 8. Also, the 8 factors and their number of levels are:

Factor 1. Size of the House, $L_1 = 9$ levels

2. Number of Stories, $L_2 = 3$ levels

3. Number of Bathrooms, $L_3 = 3$ levels

4. Presence or absence of Air-Conditioning, $L_4 = 2$ levels

5. Parking, $L_5 = 2$ levels

6. Type of Foundation, $L_6 = 2$ levels

7. Geographic Location, $L_7 = 12$ levels

8. Metropolitan Location, $L_8 = 2$ levels

The dependent variable is the price of the house and as said before is practically a continuous variable.

Determining Data Sets

Consider the total set of response data. Separate this set into 3 subsets:

- Subset D1. Complete information: the value of the dependent variable (here the price of a house) and the level of each one of the I factors are given; that is to say, an answer has been given for each question on the questionnaire.

- Subset D2. Incomplete information: the value of the dependent variable is given, but at least one of the levels of the factors was not reported.

- Subset D3. The dependent variable is not observed.

Since the dependent variable is continuous and the independent variables are discrete and since this imputation procedure requires the data to be classified in cells, imputation will be considered only for the independent variables, i.e., only for subset D2.

Classifying of the Data into Cells

Consider all $\prod_{K=1}^{I} L_K = C_1$ cells. (For the above factor levels, $C_1 = 15{,}552$). Classify the complete data of subset D_1 into these cells. Among these C_1 cells, consider the C_2 cells, $C_2 \leq C_1$, each containing "enough" observed values of the dependent variable so that a distribution can be assigned to these values. Only these C_2 cells will be considered in the imputation procedure. Therefore, the assumption is made that the incomplete questionnaires of the subset D2 can be classified appropriately using only these C_2 cells. If it happens that all C_2 cells contain enough observations so that the empirical distributions would be good enough approximations for the theoretical distributions, then the empirical distributions can be used instead of the theoretical ones.

Consider an incomplete questionnaire from subset D2. Determine from the C_2 cells those corresponding to the same level of non-missing factors which are given by the respondent. Suppose that C_3 cells are thus selected. (If the answers corresponding to factors 2 and 3 of our first example are missing, a maximum of 3 x 3 = 9 cells are isolated from the C_2 cells; these 9 cells are determined by the given levels of factors 1 and factors 4 through 8. There are nine such cells since factors 2 and 3 each have 3 levels. So here $C_3 \leq 9$.) The set of C_3 cells constitutes the possible non-empty cells from which the incomplete questionnaire may have originated.

Specifically, if all possible cells C_1 are made of boxes $B_1, B_2, B_3, \ldots, B_{15552}$, it may happen that the C_2 cells containing enough complete observations from the subset D1 are:

$$B_1, B_4, B_7, B_8, B_{11}, \ldots, B_{4000}.$$

Suppose that all possible cells, which an incomplete questionnaire of D2 could have come from, are $B_1, B_2, B_3, \ldots, B_9$. Then the C_3 cells selected from the C_2 cells will be the following 4 cells: B_1, B_4, B_7, B_8.

So here $C_3 = 4$.

The Classification of Incomplete Questionnaires into Cells

Assign costs of misclassification. Let $C(i|j)$ = cost of misclassifying an observation in cell i when in fact it is from class j; i, j = $1,2,\ldots,C_2$, $i \neq j$. Suppose the distributions considered have pdf's. (Discrete probabilities can be treated in a similar fashion.) Let $p_i(x)$ be the pdf for the ith cell. Let R_i be the region of classification in the ith cell; i.e., if $y \in R_i$, then y (the price of the house) is classified in the ith cell. It follows that the probability of correctly classifying y into the ith cell is

$$P(i|i,R) = \int_{R_i} p_i(x)dx. \tag{1}$$

The probability of misclassifying y into the jth cell when in fact it comes from the ith cell is

$$P(j|i,R) = \int_{R_j} p_i(x)dx. \tag{2}$$

If a priori probabilities q_i are known, we can calculate the conditional probability of a given observation y coming from class i. It is

$$\frac{q_i p_i(x)}{\sum_{i=1}^{C_3} q_i p_i(x)}. \tag{3}$$

The expected cost of misclassification is

$$\sum_{j=1}^{C_3} q_j \left\{ \sum_{\substack{i=1 \\ i\neq j}}^{C_3} C(i|j)P(i|j,R) \right\}. \tag{4}$$

If we classify the observation in class j, the expected cost is

$$\sum_{\substack{i=1 \\ i\neq j}}^{C_3} \frac{q_i p_i(x)C(j|i)}{\sum_{k=1}^{C_3} q_k p_k(x)}. \tag{5}$$

The expected cost is minimized by choosing j so as to minimize (5). This is equivalent to considering

$$\sum_{\substack{i=1 \\ i\neq j}}^{C_3} q_i p_i(x)C(j|i), \tag{6}$$

and choosing that j which minimizes it.

If more than one j minimizes (6), any one of them could be chosen. If a priori probabilities are unknown, we can do the following:

The conditional cost if the observation comes from cell i is

$$r(i,R) = \sum_{\substack{j=1 \\ j\neq i}}^{C_3} C(j|i)P(j|i,R), \tag{7}$$

and the classification procedure chooses that i which minimizes (7). Suppose now that we have a priori probabilities for the population, say q_j for cell j, $1 \leq q_j \leq C_3$, where

$$q_j = \frac{\text{number of observations in cell j}}{\text{number of observations in the } C_3 \text{ cells}}. \tag{8}$$

This assumption simplifies the problem and is reasonable for the two examples given above. Suppose further that the costs of misclassification are equal. Let p_j be the pdf of the jth class. Then the observation considered is classified in class j if $q_j p_j > q_i p_i$ for all $i \neq j$, $1 \leq i \leq C_3$.

This procedure minimizes the expected cost of misclassification assuming that the costs of misclassification are equal. (Anderson [1, p. 148].)

III. OPTIMAL PROPERTIES

1. As shown in T. W. Anderson [1, pp. 142-147],

the procedure $R = \{R_1, R_2, \ldots, R_{C_3}\}$ is admissible.

2. In a regression situation, we have the following:

Using the set D1 of complete data:

$$Y^{(1)} = X^{(1)}\beta + E, \qquad \hat{\beta} = (X^{(1)\prime}X^{(1)})^{-1}X^{(1)\prime}Y^{(1)}. \tag{9}$$

Using the set $D1 \cup D'2$, where $D'2$ are those questionnaires of D2 which were completed by imputation, we have

$$Y = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \end{pmatrix} = X\beta + E = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \beta + E. \tag{10}$$

Then given $X^{(1)}, X^{(2)}$,

$$\hat{\beta} = (X'X)^{-1}X'Y \text{ is unbiased} \tag{11}$$

$\Rightarrow$ it is unconditionally unbiased.

The generalized variance of $\hat{\beta}$ is

$$\frac{\sigma^2}{|X'X|} = \frac{\sigma^2}{|X^{(1)\prime}X^{(1)} + X^{(2)\prime}X^{(2)}|}. \tag{12}$$

Then the generalized variance of $\hat{\beta}$ is reduced if $X^{(1)\prime}X^{(1)}$, $X^{(2)\prime}X^{(2)}$ are positive definite matrices as shown below.

Theorem. If A^{nxn} and B^{nxn} are positive definite matrices, then

$$|A + B| > |A| + |B| > \max(|A|, |B|)$$

Proof. Use the following properties of positive definite matrices. (See Anderson, [pp. 333-341], for instance.) If A and B are positive definite matrices:

(i) their determinants are positive

(ii) there exists a non-singular matric C such that

$C'AC = I$ and $C'BC = D = \text{Diag.}\ (d_1, d_2, \ldots, d_n)$,

with $d_1, d_2, \ldots, d_n$ all positive.

Then $|A| + |B| > \max(|A|, |B|)$ holds. Next

$$|A + B| > |A| + |B| \iff |C'|\,|A + B|\,|C| > |C'|\,|A|\,|C| + |C'|\,|B|\,|C|.$$

$$\iff |C'AC + C'BC| > |C'AC| + |C'BC|$$

$$\iff |I + D| > |I| + |D|, \text{ using (ii) with } D = C'BC$$

$$\iff \prod_{i=1}^{n} (1 + d_i) > 1 + \prod_{i=1}^{n} d_i,$$

which holds since $d_1, d_2, \ldots, d_n$ are all positive.

IV. SUMMARY AND CONCLUSIONS

An imputation procedure for missing factor levels in an ANOVA situation is described. This procedure uses discriminant analysis to classify incomplete questionnaires into cells, and the classification leads to completion of these questionnaires. The procedure is admissible, and under the usual assumptions of linear models the generalized variance of $\hat{\beta}$ in $Y = X\beta + E$ is reduced.

Work is presently undertaken on an actual problem at the Bureau of the Census, where 139 out of 197 incomplete questionnaires were rendered complete by this procedure. The question of assigning distributions has not been resolved; but it has been circumvented in this problem by assuming that the square root of the dependent variable is normally distributed. Also assumed are prior probabilities and equal costs of misclassification.

In conjunction with the above, the following research is being pursued:

1. The determination of distributions for the C_2 cells (non-empty cells containing completed questionnaires) in an automatic fashion;

2. The determination by simulation of the accuracy of the imputation procedure; and

3. The determination of its usefulness by comparison of the regressions on $D1 \cup D'2$, where D1 refers to the completed questionnaires and where $D'2$ is that subset of the originally incomplete questionnaires that were later completed by imputation.

ACKNOWLEDGMENT

The authors gratefully acknowledge the helpful suggestions concerning this problem given them by Dr. Roger C. Pfaffenberger of the University of Maryland and Dr. William E. Strawderman of Rutgers University.

REFERENCES

[1] Anderson, T.W., *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, 1958.

[2] Chapman, David W., "A Survey of Non-Response Imputation Procedures," paper presented at a meeting of the Methodology Section of the Washington Statistical Society, Washington, D.C., April 14, 1976.

[3] Ferguson, Thomas S., *Mathematical Statistics: A Decision Theoric Approach*. New York: Academic Press, 1967.

[4] Hansen, M. H., Hurwitz, W. N. and Madow, W. G., *Sample Survey Methods and Theory, Vol. I.* New York: John Wiley & Sons, Inc., 1953.

[5] Lilliefors, H. W., "On the Kolmogorov-Smirnov test for Normality with Mean and Variance Unknown," *Journal of the American Statistical Association,* Vol. 62, June 1967, pp. 399-402.

[6] Nordbotten, Svein, "The Efficiency of Automatic Detection and Corrections of Errors in Individual Observations as Compared With Other Means for Improving the Quality of Statistics." *Proceedings of the 35th Session, International Statistical Institute,* Volume XLI, Book I, Tome XLI, pp. 417-435, Beograd, 1965.

[7] Szameitat, K. and Zindler, H. J., "The Reductions of Errors in Statistics by Automatic Corrections." *Proceedings of the 35th Session, International Statistical Institute,* Volume XLI, Book I, Tome XLI, Beograd, 1965.

[8] U.S. Bureau of the Census, "1960 Censuses of Population and Housing - Processing the Data." Washington, D.C., 1962.

[9] U.S. Bureau of the Census, "The Current Population Survey - A Report on Methodology." Technical paper No. 7, U.S. Government Printing Office, Washington, D.C., 1963.