THE ROBUSTNESS OF HOT-DECK AND CELL MEAN METHODS IN RETAINING POPULATION COVARIANCE STRUCTURE IN IMPUTED SAMPLES

Javaid Kaiser, Virginia Tech

Hot-deck method has been very successful in imputing missing values in large data sets. When values are missing at random, it produces unbiased estimates of population means (Hinkins, 1983; Kalton & Kasprzyk, 1983, Kaiser, 1986). However, the degree to which imputed values alter the covariance structure of the data matrix is not completely known.

The present study focussed on three issues: (a) to determine the variation in the covariance structure caused by imputation, (b) to explore the efficiency of hot-deck methods in small samples, and (c) to study the robustness of hot-deck and cell mean methods when the assumption of missing values missing at random is violated.

METHOD

Three variations of hot-deck and cell mean methods were compared using 3x3x4 factorial design. Hot-deck variations included were hot-deck sequential, hot-deck distance, and hot-deck random. A detailed description of these methods is available in Oh & Scheuren (1980). The factors studied were sample size (n = 30, 60, 120), the proportion of incomplete records in a sample (IR = 10%, 20%, 30%), and the number of missing values per records (MV = 12.5%, 25%, 37.5%, 50%). The design matrix was replicated 100 times. Although it is very uncommon to see sample sizes of 30, 60, and 120 in general public surveys, their use is very frequent when sampling businesses of a particular kind from a town, school districts in a state or school sites within a school district.

The correlation matrix given in Table 1 was used as population correlation matrix for this study. Data matrices of multivariate normal deviates were generated from this population matrix at random. The variance-covariance structure of every matrix generated was tested against the known population variance-covariance using the equation given below.

$$-2\log\lambda = pn(\log n - 1) - n\log|B\psi^{-1}| + tr\,(B\psi^{-1})$$

where p = number of variables in the matrix
n = sample size
B = sum of squares and sum of products matrix
ψ = population variance-covariance matrix

The test statistic $-2\log\lambda$ is asymptotically distributed as chi-square distribution with $p(p+1)/2$ degrees of freedom (Anderson, 1958). The data matrices that had variance-covariance structure similar to that of the population ($p > .05$) were retained for use in this study. One hundred matrices were generated for each cell of the design matrix.

The first variable of the data matrix was used as an exogenous variable to artificially create missing values missing systematically. The second and third variables were used for stratification purposes. Three categories were created on each of these variables by establishing cut-off points at -1.0 and 1.0. These categories combined resulted in a total of nine strata. The remaining nx8 submatrix was used for imputation purposes.

Once the data matrix was tested for its variance-covariance structure by the equation described earlier, missing values were created at random as per cell specification of the design matrix. Imputation methods were applied one at a time to impute these artificially created missing values. Hot-deck sequential used the observed value of the respective variable from the immediately preceding complete record as an estimate of the missing value. Hot-deck distance used the distance function to find the donor record from 10 potential donors; five above the record having missing value and five below it. Hot-deck random selected an observed value at random and used it as an estimate. The cell mean method imputed cell mean on the respective variable as an estimate of the missing value. The imputed matrices were tested again, for their variance-covariance structure against the population covariance structure using the equation described earlier at .10 and .05 levels of significance. The number of matrices that could not retain population covariance structure at a given level of significance because of imputation, were identified along with the imputation method used. The frequency of matrices

with altered covariance structure recorded for all imputation methods over all replications, at desired significance levels provided the database to determine the relative efficiency of imputation techniques.

After handling missing values missing at random, the original matrix that had been generated, was retrieved again. This time, missing values were created systematically using the following model. A high value on the exogenous variable (V1) caused the first and third variables to show missing values. The variables 5 and 8 showed missing values whenever observed value on variable 3 exceeded .4. Variables 6 and 7 showed missing values when the observed value on variable 3 was .4 or less. Once the missing values having a systematic pattern of occurrence were created, the same four imputation methods were used to impute the missing values. The imputed matrices were tested for changes in the covariance structure at .10, and .05, levels of significance. The statistic showing number of matrices that could not retain population covariance structure because of imputation was compiled for all four imputation methods over all replications.

RESULTS

Table 2 listed the proportion of matrices whose variance-covariance structure was significantly altered ($p < .05$) by the imputation method. The results indicated that all four methods were less efficient when they imputed missing values that were missing systematically than when they were occurring at random. It was also observed that all four methods yielded more than 5% matrices with altered covariance structure at .05 level of significance when incomplete records had 50% values missing on more than 20% records. It appeared that imputation methods were effective in retaining population covariance structure while imputing missing values when the incomplete records had less than 40% values missing and the pattern of occurrence of missing values was random. The cell mean method was found relatively inferior to hot-deck variations in almost all experimental conditions when values were missing systematically. There were not significant differences among hot-deck variations and cell mean method when the pattern of missing values was random. Overall, hot-deck random seemed to do better than hot-deck sequential and hot-deck distance methods.

Table 3 presents the proportion of imputed matrices that did not retain population covariance structure at .10 level of significance. It appeared that in larger samples hot-deck sequential is more effective than hot-deck random and hot-deck distance in retaining population covariance structure. Cell mean method was, again, found least desirable as it produced a very large number of matrices that failed the test of equal covariance with that of the population. Random selection of donor value seemed a better choice than a complex hot-deck distance method.

The results indicated that hot-deck variations are superior to cell mean method in retaining population covariance structure and therefore, support the use of hot-deck method as a new alternative to impute missing values in small samples. The data also revealed that the covariance of as many as 17% matrices many be affected ($P < .10$) when values are missing with a systematic pattern. Effective edit rules were found of significant importance to minimize this number. It appeared that without effective edit rules, the full strengths of the hot-deck method cannot be realized.

CONCLUSION

The results revealed that hot-deck variations are better than cell mean method in retaining population covariance structure after imputation, irrespective of the pattern of missing values. It was also observed that the covariance of more matrices was altered as a result of imputation when values were missing systematically than when they occurred at random. The data indicated that imputing more than 40% missing values per record when 20% or more records are incomplete, has a very high likelihood of changed covariance structure. Whenever imputations are made, it is suggested that the variance-covariance structure of imputed matrix be compared with the original sample matrix to gain insight about the changed covariance.

References

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. N.Y.: John Wiley & Sons, Inc.

Cox, Brenda, G. (1983). Sources and solutions for missing data in the NMCUES. Proceedings of the Section on Survey Research Methods, ASA (444-449).

David, M. & Triest, R. (1983). The CPS Hot-Deck: An evaluation using IRS records. Proceedings of the Section on Survey Research Methods, ASA (421-426).

Kaiser, Javaid. (1986). Comparison of hot-deck variations in imputing missing values. Proceedings of the Section on Survey Research Methods, ASA (653-656).

Kalton, G. & Kasprzyk, D. (1983). Imputing for missing survey responses. Proceedings of the Section on Survey Research Methods, ASA (22-31).

Oh, H. L. & Scheuren, F. J. (1980). Estimating the variance impact of missing CPS income data. Proceedings of the Section on Survey Research Methods, ASA (408-415).

Table 1

Population Correlation Matrix

	VI	V2	V3	V4	V5	V6	V7	V8
VI	1.00							
V2	.318	1.00						
V3	.468	.230	1.00					
V4	.403	.317	.305	1.00				
V5	.321	.285	.247	.227	1.00			
V6	.414	.272	.263	.322	.187	1.00		
V7	.365	.292	.297	.339	.398	.388	1.00	
V8	.413	.232	.250	.380	.441	.283	.463	1.00

Table 2

The Proportion of Sample Matrices that Could not Retain Population Covariance Structure at .05 Level

			Hot-deck							
			Sequential		Distance		RANDOM		CELL MEAN	
N	MR	MV	R	S	R	S	R	S	R	S
30	.1	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		2	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01
		3	0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.01
		4	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.01
	.2	1	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01
		2	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.02
		3	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.03
		4	0.04	0.05	0.04	0.05	0.03	0.04	0.03	0.06
	.3	1	0.01	0.03	0.01	0.02	0.01	0.03	0.01	0.04
		2	0.02	0.03	0.02	0.04	0.02	0.03	0.01	0.06
		3	0.04	0.05	0.04	0.06	0.05	0.06	0.05	0.12
		4	0.07	0.12	0.06	0.11	0.07	0.11	0.07	0.19
60	.1	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
		2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		3	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.01
		4	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.02
	.2	1	0.00	0.03	0.00	0.02	0.00	0.01	0.00	0.02
		2	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.03
		3	0.01	0.04	0.01	0.03	0.01	0.02	0.01	0.05
		4	0.03	0.05	0.03	0.04	0.01	0.03	0.03	0.09
	.3	1	0.00	0.05	0.00	0.04	0.00	0.03	0.00	0.06
		2	0.02	0.05	0.02	0.05	0.01	0.04	0.01	0.12
		3	0.04	0.10	0.04	0.09	0.03	0.08	0.02	0.26
		4	0.07	0.14	0.06	0.14	0.06	0.12	0.06	0.39
120	.1	1	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01
		2	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01
		3	0.01	0.01	0.01	0.02	0.00	0.01	0.00	0.01
		4	0.01	0.02	0.01	0.02	0.01	0.01	0.01	0.02
	.2	1	0.00	0.04	0.00	0.06	0.00	0.04	0.00	0.06
		2	0.01	0.04	0.00	0.05	0.00	0.04	0.00	0.09
		3	0.03	0.05	0.03	0.07	0.01	0.05	0.01	0.16
		4	0.05	0.09	0.04	0.10	0.02	0.08	0.05	0.23
	.3	1	0.00	0.15	0.01	0.14	0.01	0.12	0.00	0.20
		2	0.02	0.12	0.02	0.11	0.01	0.11	0.02	0.36
		3	0.05	0.18	0.06	0.18	0.02	0.17	0.04	0.61
		4	0.12	0.15	0.09	0.29	0.05	0.27	0.11	0.73

MR: Number of incomplete records
MV: Number of missing values per record
R: Random pattern of missing values
S: Systematic pattern of missing values

Table 3

The Proportion of Sample Matrices that Could not Retain Population Covariance Structure at .10 Level

			Hot-deck							
			Sequential		Distance		RANDOM		CELL MEAN	
N	MR	MV	R	S	R	S	R	S	R	S
30	.1	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	.2	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		3	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.01
		4	0.01	0.02	0.01	0.01	0.01	0.02	0.00	0.01
	.3	1	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01
		2	0.00	0.01	0.00	0.01	0.01	0.01	0.00	0.02
		3	0.01	0.02	0.01	0.02	0.02	0.02	0.02	0.06
		4	0.03	0.07	0.03	0.05	0.03	0.05	0.02	0.10
60	.1	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	.2	1	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01
		2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		3	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01
		4	0.02	0.01	0.01	0.01	0.00	0.01	0.00	0.03
	.3	1	0.00	0.01	0.00	0.02	0.00	0.02	0.00	0.02
		2	0.01	0.02	0.00	0.02	0.01	0.01	0.00	0.04
		3	0.02	0.04	0.01	0.03	0.01	0.02	0.00	0.13
		4	0.03	0.06	0.02	0.06	0.02	0.05	0.03	0.23
120	.1	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		3	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
		4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	.2	1	0.00	0.01	0.00	0.02	0.00	0.01	0.00	0.01
		2	0.01	0.01	0.00	0.02	0.00	0.00	0.00	0.02
		3	0.01	0.02	0.02	0.03	0.01	0.02	0.00	0.07
		4	0.02	0.04	0.01	0.05	0.00	0.03	0.01	0.13
	.3	1	0.00	0.06	0.00	0.06	0.00	0.05	0.00	0.10
		2	0.00	0.05	0.01	0.06	0.00	0.06	0.00	0.20
		3	0.01	0.07	0.02	0.09	0.01	0.08	0.01	0.46
		4	0.04	0.07	0.04	0.17	0.01	0.14	0.05	0.60

MR: Number of incomplete records
MV: Number of missing values per record
R: Random pattern of missing values
S: Systematic pattern of missing values