THE EFFECTIVENESS OF HOT-DECK PROCEDURES IN SMALL SAMPLES

Javaid Kaiser, University of Kansas

The hot-deck method is a simple and useful technique to impute missing values in a data matrix. It is primarily a record matching technique in which an incomplete record is compared with a complete record having similar characteristics (Ernst, 1978; Rander, 1978). The missing field in the incomplete record is imputed from the value which appears on the corresponding variable in the complete record.

The hot-deck method has been used successfully in the past by the American and Canadian Census Bureaus, the Social Security Administration, and the Internal Revenue Service. However, application of this procedure to date has been limited to large data sets. Because of the lower cost and the simplicity with which hot-deck methods may be used in real life situations it seemed advantageous to determine their usefulness when applied to small samples. The present study was an attempt to provide this information.

Purpose of the study

Hot-deck methods are typically used with large data sets. The purpose of this study was to test their robustness in small samples. Three hot-deck variations were examined to determine their relative efficiency in estimating missing values. The relative effectiveness of the three metods was determined by (1) the quality of missing value estimates produced, (2) the extent to which the estimates of population means were biased, and (3) the degree to which the population covariance structure was retained in the imputed samples. A single best procedure was identified under each criterion to help researchers in choosing the most appropriate hot-deck variation.

Hot-deck variations

The quality of estimates for missing values depends on how a donor record is selected and used in the hot-deck method. The excessive use of a single donor results in poor estimates. Several strategies have been proposed to select donors and each such strategy represents a unique hot-deck variation (Bailar and Bailar, 1978; Colledge, Johnson, Pare, and Sande, 1978). As a result, several variations of the hot-deck method have emerged.

Three variations of the hot-deck method were investigated in this study. The methods differed from one another in the way donor records were selected. In the first variation, the immediately preceding complete record was used as a donor to impute missing values to the incomplete record. This form of hot-deck is known as the sequential hot-deck method. In the second variation, all the complete and incomplete records were pooled randomly to make a hot deck. Missing fields in an incomplete record were imputed by selecting a donor at random from the complete records present in the hot deck. As recommended by Schieber (1978), each complete record was allowed to be a donor only once. Imputed records were not used as donors. In a situation where incomplete records in a particular stratum outnumbered the complete records, secondary imputation was in order and was carried out as described by Colledge et. al. (1978). The third variation considered in this study also used the hot-deck of complete and incomplete records. The nearest complete record which was not necessarily the immediately preceding record, was used as the donor. When a missing value was equidistant from the two equally elegible donors, the mean of the two donor values was imputed in the missing field. The third variation of the hot-deck was created after the suggestion of Bailar et al. (1978). In all the three variations, no edit rules were applied on incomplete records to change imputed values or the adjacent observed values.

Design and Data generation

Design:

The design of the study was a 3x4x3 factorial. The proportion of incomplete records (π) and the number of missing values in a record (ϕ) were treated as between-group factors. The three variations of the hot-deck method were the within-group factor. The proportion of incomplete records had three levels: .1, .2, and .3. The number of missing values in a record had four levels: 1, 2, 3, and 4. The total design matrix had 12 cells and each cell was replicated 500 times. All the samples used in the study, were of size 30 and had 10 variables in it. The first two variables were treated as stratification variables. The remaining eight variables were used for imputation.

Data Generation:

Ten variables having intercorrelations in the range of .19 to .47 were selected from the literature. The resulting 10x10 correlation matrix was used as a correlation matrix of a multivariate normal population. Five hundred 30x10 data matrices of multivariate normal deviates were generated from this population with the help of an IMSL computer routine. These matrices were treated as multivariate normal random samples in standard score form.

Procedure

The stratification variables were recoded from an interval scale to nominal scale in all the samples generated for this study. The cut-off points were established such that i=1 if i<=

-1.0; i=2 if -1.0 <i<= 1.0; and i=3 if i> 1.0, where i was the value on the stratification variable. This scheme resulted in a total of 9 possible strata. The values on the stratification variables for a given record represented the stratum to which the record belonged.

In order to create missing values, m records were selected randomly from the sample to represent incomplete records. The value of m was determined from the design specifications of the cell such that $m = \pi n$. On each selected record, values were randomly selected and were changed to missing value code. Random numbers generated from a uniform distribution (0,1) were used to select the records and the variables within each record to represent missing values. This procedure was repeated on every sample and for each cell of the design matrix.

The three variations of the hot-deck method were applied one after the other to impute the missing values created in the sample. The discrepancy in means between the complete and imputed samples was determined to study the distributional properties across all the replications within a given cell of the design matrix. A variance-covariance matrix was also computed for each imputed sample and was compared with the population covariance matrix using the method described by Anderson (1958). The statistic D was computed for each variation of the hot-deck to determine its relative efficiency in retaining the covariance structure of the complete sample in the imputed sample. This statistic proposed by Timm (1970) and modified by Gleason and Staelin (1975) is the root-mean-square deviation of off-diagonal elements of two covariance matrices representing imputed and complete samples, respectively. The quality of imputed values was determined by computing the statistic Q for each hot-deck variation under study. The statistic Q represented the root-mean-square standardized residuals between the true and imputed values (Gleason et al., 1975). The relative efficiency of the hot-deck variations was determined in terms of the degree to which a particular method retained the population covariance structure in the imputed samples, the overall quality of imputation, and the distribution of mean discrepancies. This procedure was repeated for every sample generated and for each cell of the design matrix.

Results

The mean discrepancy on the variables in complete samples and the samples imputed by the hot-deck random method was averaged over 500 samples and is recorded in Table 1 for all levels of π and ϕ. The standard deviation of the discrepancies in means also are listed on the second line in the same Table.

The analysis revealed no systematic change in the discrepancy of means as the values of π and ϕ increased. However, the standard deviation increased with the increase in ϕ for all levels of π. The standard deviations were also higher within higher levels of π compared to the corresponding values in lower levels of π. Though more than 50% discrepancies in means were negative at every level of ϕ and across all levels of π, they did not seem to be a function of π or ϕ values. The data also revealed that the variable means on samples imputed by the hot-deck random method converged to the population means.

The mean discrepancies on variable means between complete samples and the samples imputed by the hot-deck sequential method were computed for all levels of π and ϕ and are recorded in Table 2. The standard deviation for each discrepancy, at all levels of π and ϕ are recorded in the Table on the second line. Each value in the Table was based on 500 samples.

The data revealed no systematic trend in the means of the mean discrepancies across any level of π or ϕ. The mean discrepancy on variables between the complete and imputed samples converged to the population means. As π increased, the hot-deck sequential method produced a higher number of negative mean discrepancies. The standard deviations for the discrepancies in means increased with the increase in ϕ. The standard deviations were also higher for higher levels of π compared to the corresponding values at the lower levels of π.

The discrepancy in means between the complete sample and the sample imputed by the hot-deck distance method was averaged over 500 samples for all levels of π and ϕ and is given in Table 3. The standard deviation for these discrepancies is also recorded in the Table on the second line for each level of π and ϕ.

No systematic trend in the discrepancy of means on sample variables was found across any level of π or ϕ. The number of values having negative discrepancies were almost the same at all levels of ϕ. The discrepancy in means converged to the population means on all the variables at least to two decimal positions. The standard deviation for each mean discrepancy increased with the increase in ϕ at every level of π. The standard deviations of discrepancy in means were also higher at higher levels of π than the corresponding values at lower levels of π.

The statistics D and Q computed to compare the three hot-deck variations are given in Table 4. The D statistic revealed that the hot-deck distance method is comparatively a better imputation procedure than the other two variations at all levels of π and ϕ. This finding was also supported by the statistic Q which measured the quality of missing value estimates.

The proportion of imputed samples that could not retain the population covariance structure is given in Table 5. The proportion of such samples increased with the increase in the values of π and ϕ, for all the three hot-deck methods. The range of proportions varied from .17 to .40 .

The results revealed that when 10% of the recods in a sample were incomplete and each such record had only one missing value, 17-18% of the samples had covariances different from that of the population they represented ($p <= .05$). The number of these samples increased to 36-40% when 30% of the sample records were incomplete and each such record had missing values on 50% of

its variables. It was also revealed that the hot-deck sequential method is comparatively a better procedure than the hot-deck random and hot-deck distance methods in retaining the population covariance structure in its imputed samples.

Discussion

The data analysis revealed that the three hot-deck variations give an unbiased estimate of population means upto two decimal positions. The estimates of population means could have been better had the stratification variables been more highly correlated with the imputable variables. The median correlation between stratification and other variables, in this study, was .27. The correlation between the stratification variables was .25. However, this finding was partially supported by the literature that the hot-deck method yields unbiased estimates of population means (Bailar et al., 1978; Ernst, 1978).

Another finding that the variances of discrepancies in means increased with the increase in π and ϕ values was also confirmed by earlier studies. Based on mathematical work, Bailar et al. (1978) found that the variance estimates depend on the covariance structure of the imputed sample. He also observed that a hot-deck procedure yielded higher variances when the sample elements were selected at random which was the case in this study.

Though all estimates of population means converged to the true value upto two decimal positions irrespective of the values of π, ϕ and the hot-deck variations used, the higher variances at higher levels of π and ϕ suggested that the estimates of population means will be poorer as the proportion of incomplete records, the number of missing values in a record, or the combination of both increase.

The finding that the hot-deck distance method had the lowest root-mean-square deviation (statistic D) between the true and imputed values supported the claim of Bailar et al. (1978). He introduced this hot-deck variation and predicted it would produce better estimates than the hot-deck sequential method. Though the sample elements were selected at random, there existed some auto correlation between two adjacent values on variables that made the immediately preceding complete record a better donor than the imputed value selected at random. Therefore, the hot-deck sequential method emerged as the second best method while the hot-deck random method was the least prefered. This ranking of hot-deck variations was also supported when the three methods were compared in terms of the root-mean-square standardized residuals (statistic Q).

The characteristic of hot-deck variations to retain the population covariance structure in the imputed samples had not been tested previously. This study found that depending on the values of π and ϕ, the covariance of 18-40% of the samples was adversely affected. The three hot-deck variations therefore, were not appropriate to impute missing values if the purpose of imputation was to perform regression, cannonical, discriminant or factor analysis on the imputed sample. This was further true at higher levels of π and ϕ.

Conclusion

The results obtained indicated that there is no difference in the hot-deck random, the hot-deck sequential, and the hot-deck distance methods in estimating population means. With the increase in proportion of incomplete records, the number of missing values in a record, or both, the quality of estimates for the population means decreased due to high variance. Though all the three methods gave a large number of samples whose covariance structure was adversely affected by imputation, the hot-deck sequential method was considered comparatively a better procedure than the hot-deck random and the hot-deck distance methods. However, in measuring the overall quality of missing value estimates produced by a given hot-deck variation, the hot-deck distance method systematically performed better than the other two hot-deck methods. The results of this study caution the data analyst about the harmful effects of imputation on sample covariance.

References

Anderson, T. W. An Introduction to Multivariate Statistical Analysis. John Wiley & Sons Inc. N.Y. 1958, pp 264-267.

Bailar, J. C., and Bailar, B. A. Comparison of two procedures for imputing missing survey values. Proceedings of the American Statistical Association, 1978.

Colledge, M. J., Johnson, J. H., Pare, R., and Sande, I. G. Large scale imputation of survey data. Proceedings of the American Statistical Association, 1978.

Ernst, L. R. Weighting to adjust for partial nonresponse. Proceedings of the American Statistical Association, 1978.

Gleason, T. C., and Staelin, R. A proposal for handling missing data. Psychometrika, 40:2, 1975.

Rander, D. B. The development of statistical matching in economics. Proceedings of the American Statistical Association, 1978.

Schieber, S. J. A comparison of three alternative techniques for allocating unreported social security income on the survey of low-income aged and disabled. Proceedings of the American Statistical Association, 1978.

Timm, N. H. The estimation of variance-covariance and correlation matrices from incomplete data. Psychometrika, 35:4, 1970.

The author is grateful to Drs. Douglas Glasnap and Nona Tollefson for reviewing the manuscript.

Table 1

Means and Standard deviations of the Discrepancy
between Imputed and Complete Sample
Means for the Hot-deck
(Random) Method

π ϕ	V1	V2	V3	V4	V5	V6	V7	V8
.1 1	-0.000592	0.000955	-0.001887	0.001478	0.000796	-0.001039	-0.000825	-0.000382
	0.025166	0.029650	0.039118	0.023143	0.021635	0.027577	0.026872	0.031136
2	-0.001429	0.002020	-0.002151	-0.000511	-0.001698	-0.000867	-0.001697	-0.001327
	0.035665	0.041355	0.055469	0.039466	0.035890	0.039140	0.039536	0.037217
3	-0.001273	-0.000642	0.002288	-0.001659	0.001368	-0.000558	0.001768	-0.001362
	0.046917	0.052635	0.070431	0.046936	0.038529	0.049286	0.046893	0.040786
4	0.003379	-0.006755	0.006990	-0.000210	-0.003725	-0.003815	0.003843	0.003826
	0.055293	0.058548	0.079180	0.058209	0.049677	0.055723	0.052810	0.055910
.2 1	-0.002803	-0.003080	0.002967	0.001022	-0.001820	-0.002252	0.001604	-0.000080
	0.040331	0.039972	0.054150	0.043838	0.034394	0.038384	0.038860	0.035553
2	0.003553	0.000389	0.004415	-0.000133	-0.000143	-0.002704	-0.003692	0.000625
	0.054268	0.057551	0.077766	0.059549	0.050566	0.054721	0.054751	0.053337
3	-0.000671	0.000409	-0.006070	0.003194	-0.000519	0.003806	-0.004158	-0.003043
	0.071994	0.075633	0.097591	0.070474	0.061924	0.066862	0.069961	0.065930
4	-0.001370	-0.002803	-0.007418	-0.004841	0.002703	-0.003378	-0.006529	0.000314
	0.082284	0.083774	0.110637	0.082602	0.069718	0.078932	0.077211	0.077276
.3 1	-0.000175	0.002656	-0.007244	-0.000462	0.001987	-0.001592	0.000017	-0.000443
	0.045656	0.054250	0.069150	0.046808	0.040816	0.044400	0.045033	0.046000
2	-0.003708	-0.001548	-0.002545	0.002916	0.000357	-0.002506	-0.000607	-0.002370
	0.069773	0.071167	0.104921	0.070778	0.061493	0.063753	0.066542	0.068779
3	-0.003296	-0.009071	-0.008536	-0.001972	-0.003789	-0.003044	-0.002141	0.002602
	0.090355	0.089338	0.123328	0.085540	0.077925	0.085404	0.083492	0.086062
4	-0.001252	-0.004900	-0.007770	0.006025	-0.006226	-0.001283	-0.002119	-0.000676
	0.090615	0.102840	0.134737	0.097969	0.089361	0.099581	0.092992	0.102210

Table 2

Means and Standard deviations of the Discrepancy
between Imputed and Complete Sample
Means for the Hot-deck
Sequential Method

π	ϕ	V1	V2	V3	V4	V5	V6	V7	V8
.1	1	-0.001812	-0.000215	-0.000458	0.000143	-0.000831	-0.000496	-0.001470	-0.001664
		0.026902	0.029289	0.040617	0.026376	0.023233	0.026989	0.027127	0.029114
	2	-0.003874	0.003720	0.000634	-0.001554	-0.003012	0.000439	0.000306	-0.000991
		0.034296	0.040369	0.057825	0.038335	0.035788	0.039789	0.035541	0.037212
	3	-0.002068	-0.002542	0.000170	-0.000781	-0.001397	0.001831	0.000217	0.000360
		0.047507	0.051741	0.067864	0.048943	0.040503	0.050135	0.046855	0.042635
	4	0.000779	-0.002856	0.004642	-0.000151	0.000526	-0.005692	0.001355	0.003756
		0.056655	0.059083	0.075027	0.055940	0.049394	0.056727	0.050908	0.053262
.2	1	-0.000952	-0.002253	0.001260	0.001335	-0.002231	-0.004040	-0.001119	0.000783
		0.040549	0.038871	0.050365	0.042648	0.033722	0.041893	0.041199	0.036838
	2	0.005361	-0.001063	0.004308	-0.001989	-0.003594	0.000191	-0.001986	0.002066
		0.058123	0.056451	0.078885	0.059688	0.049991	0.051564	0.054962	0.053844
	3	-0.004154	-0.001149	-0.003001	0.002275	-0.000930	0.003587	-0.004774	-0.008410
		0.068735	0.076771	0.101920	0.073552	0.061070	0.067041	0.066880	0.067154
	4	-0.002138	0.002115	-0.004938	-0.000346	-0.002015	-0.003511	-0.005636	0.004252
		0.083715	0.080017	0.119684	0.079420	0.070086	0.083028	0.078487	0.078908
.3	1	-0.001438	-0.000455	-0.004414	-0.001091	0.000860	-0.001450	-0.000979	0.000395
		0.046855	0.052563	0.071048	0.045866	0.039849	0.047293	0.047502	0.049981
	2	-0.006491	-0.003803	-0.004014	-0.001162	-0.002722	-0.001466	0.001771	0.002543
		0.070180	0.071452	0.102052	0.072349	0.060614	0.064088	0.070184	0.066132
	3	-0.004922	-0.010722	-0.007663	-0.002255	-0.004365	-0.002742	-0.005518	-0.001669
		0.087401	0.086426	0.134095	0.091130	0.080793	0.090862	0.080068	0.083195
	4	-0.003125	-0.002586	-0.003325	0.000266	-0.010015	-0.003455	-0.003735	-0.004438
		0.096334	0.101347	0.139313	0.098722	0.090221	0.094841	0.098904	0.102352

Table 3

Means and Standard deviations of the Discrepancy between Imputed and Complete Sample Means by the Hot-deck (Distance) Method

π	ϕ	V1	V2	V3	V4	V5	V6	V7	V8
.1	1	-0.001627	-0.000629	-0.001378	0.000892	-0.000394	0.000190	-0.001256	-0.000856
		0.024819	0.026529	0.036698	0.024583	0.021718	0.026349	0.024027	0.028941
	2	-0.003170	0.000354	0.000355	-0.001598	-0.003013	0.001912	-0.000022	-0.001245
		0.032921	0.037931	0.052535	0.035748	0.032586	0.037390	0.035803	0.033432
	3	-0.000652	-0.001916	0.006146	-0.000100	0.001409	0.000042	0.001271	-0.000212
		0.044543	0.046618	0.064863	0.045402	0.037261	0.044474	0.044223	0.038434
	4	0.000818	-0.005608	0.007973	-0.000642	-0.000833	-0.006042	0.000467	0.003063
		0.052239	0.054524	0.074075	0.054184	0.046695	0.050327	0.049253	0.049334
.2	1	-0.000613	-0.002168	0.001088	0.001733	-0.000676	-0.003247	0.002199	0.000110
		0.039135	0.037432	0.047737	0.039237	0.032035	0.036659	0.036727	0.034720
	2	0.003426	-0.000986	0.001622	-0.000341	-0.000760	-0.001737	-0.001783	0.002453
		0.054515	0.054015	0.073700	0.055684	0.046680	0.049529	0.050851	0.049743
	3	-0.002875	0.002604	-0.002675	0.001550	0.000236	0.000915	-0.002517	-0.006534
		0.066192	0.070982	0.094552	0.067477	0.058253	0.063413	0.062292	0.063455
	4	0.002446	0.006509	-0.002724	-0.000401	0.000966	-0.003395	-0.004194	0.002718
		0.074018	0.077824	0.109278	0.072972	0.067962	0.077135	0.074805	0.072119
.3	1	0.000235	-0.000921	-0.004792	0.000942	0.002329	-0.000731	0.000631	-0.002032
		0.042055	0.049091	0.067468	0.042658	0.040756	0.043256	0.042603	0.044314
	2	0.004658	-0.000961	-0.007515	0.001941	-0.000762	-0.000968	0.004487	0.001744
		0.067450	0.069780	0.096951	0.064152	0.057794	0.061034	0.060955	0.063913
	3	-0.002907	-0.005587	-0.009903	0.001267	0.001017	-0.000440	-0.002020	0.076518
		0.080915	0.083494	0.122021	0.084311	0.074491	0.080092	0.073954	0.031108
	4	0.001255	-0.004628	-0.000183	0.003403	-0.005979	-0.000022	-0.002514	-0.001442
		0.088178	0.101616	0.128617	0.092835	0.089158	0.093056	0.088941	0.091680

Table 4

Statistics D and Q for the Three Hot-deck Variations

		Statistic 'D'			Statistic 'Q'		
π	ϕ	Random	Sequential distance		Random	Sequential Distance	
.1	1	0.043245	0.043039	0.040178	1.228968	1.260348	1.165273
	2	0.061862	0.060794	0.057789	1.270482	1.251678	1.175998
	3	0.074479	0.075485	0.070240	1.287173	1.301282	1.206621
	4	0.087320	0.085666	0.080632	1.316227	1.299105	1.214823
.2	1	0.063529	0.064470	0.059621	1.289484	1.302735	1.215898
	2	0.114444	0.109613	0.107150	1.313042	1.293990	1.237300
	3	0.114905	0.114780	0.109356	1.330143	1.324732	1.252357
	4	0.126874	0.126624	0.120594	1.318305	1.323048	1.237715
.3	1	0.080639	0.080067	0.077682	1.290954	1.288528	1.224505
	2	0.116736	0.117549	0.112925	1.317452	1.313378	1.250884
	3	0.142928	0.141820	0.135847	1.328718	1.328221	1.246887
	4	0.160551	0.159086	0.154688	1.321899	1.319008	1.256616

Table 5

The Proportion of Imputed Samples whose Covariance Structure Significantly differed from that of Population at .05 Level

	Number of Missing Values											
	1			2			3			4		
π	R	S	D	R	S	D	R	S	D	R	S	D
.1	.182	.180	.172	.184	.180	.186	.216	.188	.182	.234	.224	.250
.2	.184	.170	.178	.194	.206	.218	.248	.238	.240	.322	.248	.306
.3	.198	.182	.202	.246	.212	.226	.312	.268	.278	.404	.362	.376