DISCUSSION

Roderick J.A.Little, Datametrics Research, Inc.

I wish to thank the organiser and authors for the chance to discuss two well written papers, and for the luxury of receiving them in time for the meeting.

1. General Remarks

Before commenting on specifics, it might be worthwhile distinguishing the following approaches to handling item non-response in sample survey data:

a) Discard units with data missing on some items, and treat non-response as a form of subsampling of the originally sampled items.

b) Impute or substitute values for the missing items and proceed as if the imputed values were the true values.

c) Leave the incomplete data matrix as it stands. When computing estimates from the data, use an efficient method of estimation (for example, maximum likelihood), under an assumed model for the item values. For a recent review of this approach, see Little (1981).

d) Impute a set of k>1 values for each missing value (multiple imputation). These sets can be used to derive nearly efficient estimates of population quantities, together with valid estimates of variance (Rubin, 1978; Herzog and Rubin, 1981).

The papers we have heard today primarily concern strategy b), although some material is devoted to a) and d). Strategies a) and b) are relatively naive, although they may be acceptable for small amounts of missing data. Strategy a) has the disadvantage that it is vulnerable to bias in the survey estimates, arising from the restriction to responding units. Strategy b) may succeed in limiting the bias of estimates, but leads to underestimates of the variance of sample estimates, since these variances are estimated without taking into account non-response. Strategies c) and d) have the potential of avoiding both these problems, but require some expertise to implement.

2. The paper by Santos.

In his interesting paper, Mr. Santos distinguishes between the problem of missing values in estimating means of variables, and in estimating functions of the covariance matrix of a set of variables, such as covariances or slopes. I was a little disappointed in the restriction to bivariate relationships, since in most analyses of multivariate data partial rather than simple regression coefficients are of interest. Also, covariances between pairs of variables seem to me to be of limited analytical interest. Perhaps the most pertinent case covered by the paper is the slope B_1 when x is a binary variable, which can be interpreted as the difference between the means of y for the two categories of x.

The basic message of the paper is that following strategy b), with missing values replaced by respondent means or values from a hot deck procedure, often leads to bias in estimates of variances, covariances and simple regression coefficients. I contend that unbiased estimates are easier to obtain than the author implies. For example, unbiased estimates of the covariance S_{xy} and the slope B_1 can be obtained by regression imputation, provided care is taken in the choice of variables included as regressors for predicting missing y values. The following choices can be distinguished:

i) Regress y on the constant term only. That is, impute the grand mean of y. This leads to Santos's (GM) method.

ii) Regress y on z, treated as categorical (CM) or as continuous (RG).

iii) Regress y on x, treated as continuous. This can be shown to be equivalent to discarding incomplete units (DN).

iv) Regress y on x, treated as continuous, and z, treated as continuous or categorical. This method is not considered by Santos.

As Santos demonstrates, the variable x must be included in the regression to avoid bias in sample estimates of S_{xy} and B_1. Thus choices i) and ii) lead to biased estimates. Choices iii) and iv) lead to unbiased estimates of S_{xy} and B_1, under certain assumptions about the mechanisms leading to missing data.

Santos notes that iii) leads to unbiased estimates when the probability of response does not depend on the values of x, y or z. If we make the model assumption that the regression of y on x in the population is linear, then this condition can be weakened. For unbiasedness it is then only necessary that for fixed x, the probability of response does not depend on values of y or z. In other words, x-values in the sample do not need to be a random sample of the x-values in the population.

The choice iv) is superior to iii) in two respects. Firstly, it is more efficient in cases when z and x jointly predict y better than x alone. More significantly, it is unbiased when the probability of response depends on the value of z. If the model assumptions underlying the analysis are correct, then it is unbiased provided the probability of response for fixed x and z does not depend on y. This condition allows the probability of response to depend on the values of z and x. More formal conditions for unbiasedness can be formulated, using the theory of Rubin (1976).

$O(n^{-1})$ unbiased estimates of the sample variance can also be derived with negligible additional computations. For example, the bias in the sample variance in equation (6.4) is easily removed by adding M times the residual variance of y given z to the unadjusted sample variance. This estimate can also be improved by including x as a regressor variable when predicting missing y's.

A final remark on the paper by Mr. Santos concerns the illustration using S.I.P. data. It is certainly valuable to compare methods on real data. However it is important to bear in mind that, although the patterns of missing data created were designed to match the patterns actually encountered, this does not ensure that the simulations are realistic. The simulations are based on an assumption that the missing data are missing at random (in the terminology of Rubin, 1976), whereas in practice the respondents and non-respondents (and in particular, the slopes B_1 for these groups) may differ after adjusting for covariates such as x and z. The validity of the missing at random assumption cannot be addressed by the data. Thus the only alternative is to attempt to evaluate the sensitivity of final estimates to plausible departures from this assumption. Some alternative approaches to this exercise are discussed in Rubin (1977, 1978) and Little (1981).

3. The paper by Kalton and Kish

In the discussion of the paper by Mr Santos, I have concentrated on methods which impute means, perhaps conditional on covariates, although the expressions for bias in estimates of B_1 and S_{xy} carry over to hot deck analogs. In the paper by Professors Kalton and Kish, the emphasis is on hot deck procedures. These methods are popular among practitioners, perhaps because of ease of implementation, although recent applications like the CPS hot deck are computationally quite elaborate. They also have the advantage over mean imputation of preserving the distribution of the imputed variable y, which is a particularly useful property when y is grouped into categories for crosstabulation.

The penalty of hot deck imputation is an increase in variance in the estimates. The main message of the Kalton and Kish paper is that the selection of m imputed values of y from a set of r candidates (perhaps in some restricted subclass c of the sample) is an unusual form of sampling, since the y-values of all the r candidates are known. Hence the candidates can be stratified by y before selecting values for imputation. This minimizes the additional variance from hot deck imputation, whilst preserving the distribution of y in the imputations.

The idea is ingenious, and seems to differ from forms of stratification discussed in the existing literature, which are directed at forming subclasses c defining a set of candidates for each value to be imputed. I might add that the latter seems to me a more important problem, since it relates to minimizing bias rather than variance. (Incidentally, I hope the authors include more references to the hot deck literature in the final version of their paper.)

Professors Kalton and Kish also mention an alternative way of reducing imputation variance, namely by performing multiple imputations and then averaging the estimates from each imputed data set. As noted above, multiple imputation has one important advantage over single imputation methods, namely that valid variance estimates can be obtained. In his discussions of the technique cited above, Rubin views the latter issue as more important than the question of variance reduction. In fact, he chooses the least efficient form of sampling, simple random with replacement, to select imputed values from the r candidates.

Kalton and Kish mention the possibility of combining multiple imputation with one of their efficient forms of sampling. I do not think this is wise, since I think it will destroy the validity of the variance estimates from multiple imputation with negligible gains of efficiency. This point may be clarified by the following simple examples. Suppose that in subclass c there are ten candidate y-values, and ten values to be imputed. Any form of random sampling without replacement yields the same set of imputed values, and hence the same mean of observed and imputed values for subclass c. As a result the added variance from the imputations is zero, but the variance estimate is too low. To take a less extreme case, suppose that five values are to be selected from ten candidates, and these values are selected by systematic sampling from the ordered list of ten y-values. If two sets of imputations are carried out, there is a 50 percent chance of obtaining the same set of imputations for both draws, again leading to a poor variance estimate.

Rubin deliberately chooses random sampling with replacement to provide unbiased variance estimates. The inefficiency introduced is negligible when two or more imputations are carried out, as his calculations have shown.

References

Herzog, T. and Rubin, D.B. (1981). Using Multiple Imputations to Handle Non-Response in Sample Surveys. Chapter for Non-Response in Sample Surveys: Theory of Current Practice, prepared by Panel on Incomplete Data, National Academy of the Sciences, Washington, D.C.

Little, R.J.A. (1981). Models for Non-Response in Sample Surveys. To be published in J.A.S.A.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika 63, 581-592.

---------- (1977). Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys. J.A.S.A. 72, 538-543.

---------- (1978). Multiple Imputations in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse (with discussion). In Imputation and Editing of Faulty or Missing Survey Data, U.S. Social Security Administration and Bureau of the Census, 1-9.