MORE ON IMPUTING VERSUS DELETING WHEN ESTIMATING SCALE SCORES

C. H. Proctor, North Carolina State University

Introduction

In an earlier paper [1] two simple solutions (deleting the case or imputing the response) to the problem posed by item non-response were compared and it was found that only when the proportion of scale items (answered among those presented) is below one half, should one delete the case. Otherwise, the results suggested it was more efficient to impute the missing item scores. The formulation in that paper was abbreviated and I have since had some opportunity to re-think the problem. It now appears that one should logically consider a number of other situations on either side of the one presented there. The case of imputation considered in that early paper is now seen to be of theoretical but not of practical interest.

It seems that one source of my own confusion with the concepts of the earlier paper was the use of the same notation in representing an error term in the model for data generation in the initial survey as for representing the imputation procedure. For example, e_{ij} represented a random person-item measurement error deviation as well as a random person-item discrepancy in the imputed value. An attempt will be made here to use the error terms b_i and d_{ij} to represent discrepancies of the imputation procedure. It is nonetheless still true that the distribution of b_i will be taken to be that of the true score a_i, and that of d_{ij} is that of e_{ij}. This is a key simplifying assumption that permits the approach to reach its definitive results. The fact that it is a realistic assumption awaits empirical verification. The following recapitulates our notation.

The response of person i to item k will be denoted as X_{ik}. For example, the values of X_{ik} may be the integers in the range 1 to 5 used to score the strongly disagree-disagree-undecided-agree-strongly agree response format. The items will be taken as roughly equivalent, as having the same variance and as being equally intercorrelated. We will also make liberal use of the normality assumption as in classical test theory. Any item response may be considered as the sum of four quantities that will be written: μ, a population average; λ_k, an item fixed effect for $k = 1,2,\ldots; K$ with $\Sigma\lambda_k = 0$; a_i, a random person effect; and e_{ik}, a random person-item measurement error effect. Both $E(a_i) = 0$ and $E(e_{ik}) = 0$ with $\sigma_a^2 = E(a_i^2)$ and $\sigma_e^2 = E(e_{ik}^2)$ over the population of persons and of their item responses; while the a_i and e_{ik} are taken to be independent of one another. The error terms b_i and d_{ij} entirely parallel a_i and e_{ij} and refer to the imputation process.

The objective of the analysis is an archtypical one of detecting in a subgroup of n persons some departure in their mean of, say μ_1, from the mean, say μ, of a much larger general population of N persons. We suppose that only one person of the n, the n^{th} say, shows non-response and that is to items PK + 1, PK + 2, ..., K while he answers items 1, 2, ..., PK. Possible values of P are 0, 1/K, 2/K, ..., 1, while Q = 1 - P. If case n's data are deleted, so to speak, then the test of $\mu_1 - \mu = 0$ will be made only with data from the n-1 completed cases. If some imputation procedure is followed for the QK missing item responses then the question arises of how to reflect or model the precision of the resulting data.

Before entering into a more detailed discussion of the imputation procedures it may be instructive to imagine ourselves facing the data in the literal and specialized experimental setting or hypothesis testing situation being assumed. There is complete information on the very large population ($N \gg n$) so that estimates of μ, the λ_j for j = 1, 2, ... K, σ_a^2 and σ_e^2 are essentially all known constants. We are then furnished with data from the subpopulation as nK + PK item responses. After using knowledge of the λ_j they will all have the unknown μ_1 as mean. They also have a known covariance structure that can be exploited to derive a linear unbiased minimum variance estimator of μ_1, say $\hat{\mu}_1$.

The test statistic

$$T_{OPT} = \hat{\mu}_1 - \mu$$

has mean $\mu_1 - \mu = \Delta$ say and variance

$$V(T_{OPT}) = \left[(n-1)/(\sigma_a^2 + \sigma_e^2/K) + (\sigma_a^2 + \sigma_e^2/PK)^{-1}\right]^{-1}.$$

This test statistic T_{OPT} is "better" than anything else to be considered later on because its variance is smaller and its non-centrality is larger. In a world of known variances and normal distributions the merit of T_{OPT}, or any such suggested test statistic, can be summarized most conveniently in its squared coefficient of variation or $\Delta^2/V(T_{OPT})$ - the larger the better. The ratio of one such quantity to another will be referred to as an "efficiency" and written EFF.

For example, deleting the n^{th} case leaves us with the mean of n-1 observations or the test statistic

$$T_D = \bar{X}_{(n-1)} - \mu .$$

$E(T_D) = \Delta$ and $V(T_D) = (\sigma_a^2 + \sigma_e^2/K)/(n-1)$, so that

$$EFF(D/OPT) = [\Delta^2/V(T_D)]/[\Delta^2/V(T_{OPT})]$$

$$= (n-1)/\left[n - \frac{Q(1-\rho)}{1-\rho+PK\rho}\right] ,$$

where $\rho = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$ is the average inter-item correlation coefficient. It may be more familiar to use in this formula the reliability r_{xx}, where $r_{xx} = K\rho/[1 + (K-1)\rho]$, and write

$$EFF(D/OPT) = (n-1)/\left[n - \frac{Q - Qr_{xx}}{1 - Qr_{xx}}\right] .$$

If the n^{th} case answers all items, that is P = 1 while Q = 0, then deleting is a very wasteful procedure as shown by the ratio EFF(D/OPT) = (n-1)/n. But when P = 0 and Q = 1 there is no information at the n^{th} case and so ignoring it will result in

no loss as reflected by EFF(D/OPT) = (n-1)/(n-1).

A certain artificiality attaches to these results since they suppose knowledge of usually unknown variances. Ordinarily one would have to use a studentized ratio as test statistic. Since we will persist with the artificiality it may be advisable to express here our **belief that the use of estimates of the variances should** not upset the findings very much. Another consideration in this vein is the distinctive nature of "deletion" as compared to "substituting the subgroup mean." Although putting the subgroup mean in place of the n^{th} person's responses would yield the same estimate of μ_1 as deleting him, the estimate of variance would be reduced when carrying along the extra observation located at the center of the other values as compared to looking only at n-1 observations. We do have in mind deleting and will maintain this terminology.

Efficiencies of Deleting Relative to Various Methods of Imputing

If the missing item responses had been obtained they would be represented as:

$$X_{n,PK+1} = \mu_1 + \lambda_{PK+1} + a_n + e_{n,PK+1}$$
$$\vdots \qquad\qquad \vdots$$
$$X_{n,K} = \mu_1 + \lambda_K + a_n + e_{n,K} .$$

In representing imputations we will suggest a somewhat similar format, however, substituting μ for μ_1 in recognition that **losing** the distinctive nature of the subgroup is part of the price one must pay for not having original data. The other features of four viewpoints are the following:

$$X'_{n,PK+1} = \mu + \lambda_{PK+1}$$
$$\vdots \qquad\qquad \vdots$$
(VI) $$X'_{n,K} = \mu + \lambda_K$$

$$X''_{n,PK+1} = \mu + \lambda_{PK+1} + d_{n,PK+1}$$
$$\vdots \qquad\qquad \vdots$$
(V2) $$X''_{n,K} = \mu + \lambda_K + d_{n,K}$$

$$X'''_{n,PK+1} = \mu + \lambda_{PK+1} + b_n + d_{n,PK+1}$$
$$\vdots \qquad\qquad \vdots$$
(V3) $$X'''_{n,K} = \mu + \lambda_K + b_n + d_{n,K}$$

$$X''''_{n,PK+1} = \mu + \lambda_{PK+1} + b_{n+1} + d_{n,PK+1}$$
$$\vdots \qquad\qquad \vdots$$
(V4) $$X''''_{n,K} = \mu + \lambda_K + b_{n+1} + d_{n,K} \quad .$$

As mentioned earlier, the distributions of d's and of b's are taken to be governed by the corresponding e's and a's. In particular, b_n is the same as a_n whereas b_{n+1} is an uncorrelated person ($n+1^{st}$) effect. Thus V4 may represent "cold deck" imputation while V3 is a rather extreme form of "hot deck" imputation - perhaps "red hot deck" would be appropriate terminology. Only viewpoint V1 among the four does not exhibit the realism of measurement errors or person effects and **is thus closest to "zero order" imputation.**

In all four cases of imputation we suppose that the test statistic is the simple average over all K item responses of all n cases minus μ. Thus its expectation, $E(\bar{X}-\mu)$, will be $[(n-1+P)\mu_1 + Q\mu - \mu]/n = (n-Q)\Delta/n$ in all four cases. The variances differ and so the resulting efficiencies are found to be:

$$EFF(D/V1) = [n-Q(1 + r_{xx} - Qr_{xx})](n-1)/(n-Q)^2$$

$$EFF(D/V2) = [n-Qr_{xx}\ (2-Q)](n-1)/(n-Q)^2$$

$$EFF(D/V3) = n(n-1)/(n-Q)^2$$

$$EFF(D/V4) = [n-2\ r_{xx}Q(1-Q)](n-1)/(n-Q)^2 \quad .$$

In the case of Q = 0 all four efficiencies agree on (n-1)/n . This says simply that deleting a perfectly good case, one of n, results in a relative efficiency of (n-1) parts in n. The situation at the other extreme, for Q = 1, differs among the viewpoints. V3 and V4 arrive at n/(n-1) which means that efficiency crosses a breakeven point between Q = 0 and Q = 1 which is to say that for some intermediate value of Q, say QBE for breakeven, it would be better to delete the case if the observed Q exceeded QBE while imputing is the preferred course if Q is less than QBE. This is a realistic performance for an otherwise theoretical model.

Efficiency for V2 also exceeds 1 for Q = 1, being $(n-r_{xx})/(n-1)$ **in fact. This shows that for** some larger values of Q it would be better to delete than to impute, although when r_{xx} is near one such a breakeven value may also be near QBE = 1 . It is under viewpoint V1 that there is no breakeven point and the result is that **one should always impute.** It is just as good to provide all K item responses for a missing case as it is to delete it. This "unrealistic" result stems from our assumptions about perfect knowledge of the item parameters and about the correctness of the model.

In case V3 the breakeven point is at

$$QBE3 = 1/2 + 1/8n + 1/16n^2 + \ldots ,$$

or just a shade above .5 . It is also possible to get a somewhat clean looking result for QBE under V4 when n is large, as

$$QBE4 = 1/2 + r_{xx}/4 - r_{xx}^3/16 + r_{xx}^5/32 - \ldots .$$

The breakeven point is nearly 1 under V2 which suggests that if <u>only</u> measurement errors are affecting the imputations then one should almost always impute. When one makes a type of imputation that even adds the person-to-person component of variance then QBE is lowered (that is, our tolerance of item non response is lowered) but even so it does not go below half of the items. From numerical inspection of a number of particular cases the conclusion is that QBE is somewhat over half of the items.

Thus I would recommend that as part of the data cleaning routine some imputation be used for item non-response if over half the items were answered.

Discussion

I find these results most encouraging to my current practices in advising researchers but rather curious from a philosophical standpoint. For example, no data were cited and none is needed but does that imply that the findings are always true? The answer is a qualified, Yes. Two basic suppositions, one on the magnitude of the loss in sensitivity and the other on the sizes of the variances involved, are so eminently reasonable that if not true their falsity should be immediately sensed. The lack of sensitivity is put at $\mu_1 - \mu$ and this will be false only if the occurrence of item non-response is perversely tied to a level uncharacteristic of the subgroup effect under test. Such effects have been noted for case non-response and require direct and special attention. If such were the case for item non-response it must be attacked directly as well. But in the present work it is our explicit assumption as well as practical judgement that μ_1 is governing our respondents from the subgroup and that $\mu_1 - \mu$ fairly represents the cost of using information from the larger population.

The other basic supposition underlies our use of error terms, with the same sized variance components as the observation, to represent imputations. Not only does this also seem to be an eminently realistic assumption but we have used a variety of imputation models, just as in practice a variety of methods of imputation are used, and found our practical conclusions to be robust.

A final feature of apparent artificiality in our assumptions should at least be mentioned again, that is, confining consideration to detecting a shift in subgroup mean with variances known when in practice most (of my) clients are doing regression analyses. I do feel there may arise some special problems, and I've even experienced them, with wholesale imputation of some single central value. This can distort estimates of variances and of covariances although it may improve estimation of the mean. Our proviso is that the kinds of imputation being considered here must not damage covariance estimation. With that proviso the use of studentization, that is in practice using estimates rather than known standard errors, will leave the results on relative efficiencies unchanged.

References

[1] Proctor, C. H. "What Proportion of Item Non-Response Should Lead to Deleting the Case?", Proceedings, American Statistical Association, Social Statistics Section, 1974, pp. 404-408.