Journal of Official Statistics, Vol. 20, No. 3, 2004, pp. 445–449

Discussion

Keith Rust[1]

It gives me great pleasure to be able to make some remarks at this, the Twelfth Annual Hansen Lecture, and I thank Nancy Kirkendall and the organizing committee for giving me this opportunity.

I worked closely with Morris Hansen during the last four years of his life, in the late 1980s. We worked together on the sample design and estimation for the National Assessment of Educational Progress (NAEP). This was a wonderful experience for me, and I learned much from my association with Morris on this project. Although perhaps not enough apparently, as I am still working on the NAEP, evidently not yet able to get the design and estimation correct, after almost 17 years. At the time that I worked with Morris on the project, NAEP had a technical advisory committee that included many noted statisticians and psychometricians. Among these was John Tukey, one of the founders of the project in around 1970. The spirited but always enlightening discussions between John and Morris on statistical philosophy and practice were as entertaining as they were educational.

One of the hallmarks of Morris's career was his efforts to ensure that research be used to inform survey practice. I believe that Morris's view was that such research needed to be rigorously designed and carried out, with a clear direction towards enhancing practice. This was evident in his own work, and I believe that he encouraged it in others. One who has clearly applied this approach to his own work is Paul Biemer. I first became personally acquainted with Paul some years before I met Morris. Although I have never had the opportunity to collaborate with Paul, I have followed his work with interest, and it gives me great pleasure to discuss his topic for today's presentation.

Paul's presentation reminds us of some of the history of methods to measure and reduce the effect of response error on survey results, specifically the contribution of Hansen, Hurwitz, and Pritzker (1964). Paul has then shown us some of the developments in this area since that time, and indicates future directions.

A major message that I took from Paul's presentation today was to avoid oversimplifying the consideration of response error, by attempting to rely on a single index – the Index of Inconsistency – introduced by Hansen, Hurwitz, and Pritzker. It is invaluable to have an index that can provide a comparative measure of the effects of response variance on survey estimates. Equally valuable is to have an understanding of the model that gives rise to that measure, and hence of its potential drawbacks and limitations if leaned on too heavily.

[1] Westat, Inc., 1650 Research Boulevard, Rockville, MD 20850-3195, U.S.A. Email: rustk1@westat.com

Paul gives several examples that show that the Index of Inconsistency tends to average over too much, and can obscure important differences in different components of the interview process. I have another concern to add to those raised by Paul.

It seems to me that focus on the Index of Inconsistency can tend to result in too much emphasis on response error as a source of variance in estimation, relative to its potential as a source of bias. The numerator of the index measures the amount of estimator variance that is contributed by response error. The denominator includes this, plus the amount of estimator variance that is contributed by sampling variance. However, when I consider the likely nature of response error, it seems to me that in most cases its impact on the bias of estimation is likely to outweigh its effect on the variance. Because the response variance is divided by the sample size in its contribution to total variance and thus mean squared error, whereas the effect of response bias is not, it seems likely that, apart from the situation of a subgroup with a very small sample size, or a very unusual circumstance, the bias is the more important consideration, especially for categorical response variables. This was apparent to me from a consideration of the Classification Probability Model that Paul described. Let me try to illustrate my concern with a hypothetical example.

Paul's article shows that bias of p as an estimator of π, the true population proportion, is equal to $-\theta\pi + \phi(1 - \pi)$ and hence p is unbiased only if $\theta\pi = \phi(1 - \pi)$, where $\theta = P(y = 0|\mu = 1)$ and $\phi = P(y = 1|\mu = 0)$ are the respective probabilities that the respondent gives the incorrect answer, given the two different possible true values.

One tends to think that "there won't be any problem with bias as long as the respondent's answers average out at the respondent's true value." But with the true value being zero or one, and the possible responses being zero or one, this can only happen if the respondent never makes a mistake. One might also think that there would be no bias if $\theta = \phi$, but the article shows that a different condition must apply. This condition states that the relationship between θ and ϕ must depend on π for unbiasedness to occur. This seems untenable in most cases.

Thus

$$MSE(p) = \frac{S_1^2}{n} + \frac{S_2^2}{n} + B^2$$

$$B = \phi(1 - \pi) - \theta\pi$$

$$I = \frac{S_2^2}{S_1^2 + S_2^2}$$

But consider B^2 relative to S_2^2/n, the components of MSE that are due to response error. Suppose that $\pi = 0.5$; $\theta = 0.02$; $\phi = 0.1$. This does not seem at first like a severe case of response error – those who have the condition of interest report that they do not 2 percent of the time, and those who do not have the condition report that they do 10 percent of the time.

Using Equations (13), (14), and (15) from Biemer's article, this gives

$$S_1^2 = 0.1936, \quad S_2^2 = 0.0548, \quad I = 0.22.$$

Thus we see that response variance contributes 22 percent of the total variance, which is in fact rather large. We also get $B = 0.04$, so that $B^2 = 0.0016$. Thus $B^2/S_2^2 = 0.029$.

This means that if the sample size is 100, the response bias contributes 2.9 times as much to the MSE as does the response variance. For a sample size of 1,000, the squared bias is 29 times the response variance, and contributes 87 percent of the total MSE. Thus I think that Paul's emphasis on using moderately complex models to describe the effects of response error, and focusing attention on a number of model parameters, is very well founded, and his interesting examples illustrate how this can be done usefully.

One of the difficulties that Paul points out must be addressed in using more realistic models of the response process is that the models can be easily under-identified. This tends to result in reliance on a series of assumptions. Often these assumptions are unlikely to be fulfilled, or else this objection is overcome by imposing further conditions on the model.

One particularly thorny issue arises in addressing the fact that any "reinterview" questions must meet the assumptions of parallelism and local dependence. Stated briefly and over-simply, this means that the question must be posed to the respondent exactly the same way for the reinterview as for the initial interview, yet the respondent's answer on the second occasion must not be influenced by what he or she reported the first time. This is just not realistic in my view, and Paul makes this point also.

I suggest that one approach that might prove promising for dealing with this is to regard the set of true values for individuals as being arrayed on an arbitrary continuous unobservable scale, rather than being a set of discrete states with known values, albeit still unobservable. Consider Paul's example of the 1998 Census Dress Rehearsal and the answer to the implied question: "Are you multiracial?" Paul describes the circumstances where the PES identified far fewer individuals as being multiracial than did the Census. Clearly the assumption of parallelism is violated here, for reasons not clearly understood. The approach that Paul describes can be paraphrased as being one in which every person is either multiracial or not, and our task is to estimate what proportion are multiracial in the face of respondents who refuse to answer the question directly for us. The alternative that I would like to consider is that there is an underlying continuum of "multiracialism," and that every individual lies somewhere along the continuum. Of course then one cannot answer the question: "What proportion of the population is multiracial?", since there is no state of being multiracial. However, one can define a cut point on the continuum (arbitrarily, but perhaps norm-referenced) above which an individual would be classified as multiracial. And one can say whether some parts of the country are more "multiracial" than others, for example.

By taking this approach, one can do away with the need for parallelism. Local independence is still required, but that is much easier to attain if parallelism is not needed. It can also be achieved in a single interview, so that a reinterview is not needed. One must find different questions on the same topic, such as Paul described in his marijuana use example.

The application with which I am familiar where this approach is taken involves what is known in psychometrics as Item Response Theory (IRT) (Adams and Wu 2002; Mislevy, Johnson, and Muraki 1992). However, this is essentially the same as the latent class (LC) approach that Paul describes, except that X, the true value, is distributed on a continuum. As Paul describes, the LC models also do away with the need for local independence. The trade-off that I see is that IRT requires fewer parameters and is

therefore more easily identified. The price for this is having to interpret that underlying continuum.

One can also use the item response theory approach, together with multiple imputation, through a Bayesian framework, to remove the bias due to response "error," measure the magnitude of the response variance, and incorporate response variance in the overall variance estimate.

The latent class model takes the form:

$$\theta_{ij} = P(A_i = j | x = 1 - j)$$

where i denotes questions and j, which takes values 0 and 1, denotes the possible values for both the true value and the response. This requires the estimation of parameters for each combination of i and j.

One form of item response theory model takes the form:

$$P(A_i = j|x) = \frac{e^{(x+\lambda_i)}}{1 + e^{(x+\lambda_i)}}$$

This only requires parameters for each value of i.

In the case of the Census Dress Rehearsal example, if A corresponds to the census question on race (with $j = 1$ denoting "multiracial") then λ is clearly larger than for the corresponding question on the PES. That is, for a given degree of "multiracialness," a respondent is more likely to classify himself or herself as multiracial in the census than in the PES.

Maximum likelihood methods can be used to estimate the values of λ. One can then use a Bayesian approach to estimate a posterior distribution for X for each individual in the sample. MLEs for the X value can be obtained, but using these to make inferences about population statistics (the population mean of X for various subgroups, for example) results in biased estimates. The procedure is to assume that X has an underlying distribution (e.g., normal), conditional on the value of various "background" characteristics, $\mathbf{Y}$. That is:

$$X \sim Normal(\mathbf{Y}^T\gamma, \sigma^2)$$

and we denote the pdf of X|$\mathbf{Y}$ as $g(X|\mathbf{Y})$.

In the case of the multiracial response in the Census, the vector $\mathbf{Y}$ might include the response to questions on Hispanicity, region of the country, gender, and age, for example. The marginal distribution for $\mathbf{A}$, $f(\mathbf{A})$, can be obtained by integrating $f(\mathbf{A}|X).g(X|\mathbf{Y}).k(\mathbf{Y})$ across X and $\mathbf{Y}$. This is then used to obtain MLEs for γ, σ, and the λ's. Applying a Bayesian approach, an individual's posterior distribution for X, given his or her response to the relevant survey questions, $\mathbf{A}$, and other characteristics $\mathbf{Y}$, is given as:

$H(X|\mathbf{A}, \mathbf{Y}) \propto f(\mathbf{A}|X).g(X|\mathbf{Y})$, with the appropriate MLEs inserted for the parameters.

Multiple (M, e.g., five) draws are made from this posterior distribution for each respondent. The estimate of a parameter of interest is made by averaging the estimates derived from each of the sets of plausible values:

$$\hat{y} = \frac{1}{m}\sum_{m=1}^{m} \hat{y}_m$$

The response variance for $\hat{y}$ is then estimated as:

$$\hat{v}_r(\hat{y}) = \left(1 + \frac{1}{m}\right)\frac{1}{(m-1)}\sum_{m=1}^{m}(\hat{y}_m - \hat{y})^2$$

This approach has been implemented effectively in the National Assessment of Educational Progress for almost twenty years, with the underlying continuum being student ability in a domain such as mathematics (see Mislevey, Johnson, and Muraki 1992).

In conclusion, I found this article very enlightening and thought provoking. I believe that Paul has presented a foundation for future developments on the topic of response error and its estimation, control, and reduction. I look forward to seeing Paul's future contributions in this area.

References

Adams, R. and Wu, M. (eds) (2002). PISA 2000 Technical Report. Organisation for Economic Co-operation and Development: Paris.

Hansen, M., Hurwitz, W.N., and Pritzker, L. (1964). The Estimation and Interpretation of Gross Differences and the Simple Response Variance. In C.R. Rao (ed.). Contributions to Statistics. Presented to P.C. Mahalanobis on the occasion of his 70th birthday, Calcutta: Statistical Publishing Society.

Mislevy, R.J., Johnson, E.G., and Muraki, E. (1992). Scaling Procedures in NAEP. Journal of Educational Statistics, 17, 131–154.

Received February 2004