AN EXTENSION OF THE RESPONSE ERROR MODEL FOR APPLICATION TO COMPLEX ESTIMATORS BASED ON DIFFERENT SAMPLES

Gad Nathan, Hebrew University, Jerusalem and Central Bureau of Statistics, Israel*

0. Introduction and Summary: The basic response error model developed by the U.S. Bureau of the Census [5] has been generalized to the multivariate case by Koch [4]. The possibilities of application of the model to complex sample designs and to complex estimators are, however, limited by the requirement that all components of the vector random variable are measured for the same sample. This is not necessarily the case for many applications, such as difference, ratio or regression estimation, or panel-type sample designs, where component estimators may be based on different samples (e.g. sub-samples, partially matched samples, etc.). In the following, the basic multivariate response error model is extended to cover the case where each component of the vector variable is measured in a possibly different sample, all selected from the same finite population. The extended model is then applied to difference and ratio estimation in different response error structure situations and to the case of sampling on two occasions.

1. The Model: The formulation of the model follows that of Koch [4]. For each unit in the population, $i=1,\ldots,N$, let there be defined a p-component vector random variable, $\underset{\sim}{Y}_{it}$, by:

$\underset{\sim}{Y}_{it}'=(Y_{it}^{(1)},Y_{it}^{(2)},\ldots,Y_{it}^{(p)})$, where t indexes the sequence of repeated trials.

Any given sample is defined by the indicator random variable:

$$U_i^{(j)} = \begin{cases} 1 : \text{if unit } i \text{ is in the sample for the } j\text{-th component} \\ 0 : \text{otherwise.} \end{cases}$$

So as to simplify the presentation, simple random sampling without replacement will be assumed for each of the component samples. Thus, if n_j is the sample size for component j, then:

$E\{U_i^{(j)}\}=n_j/N$ and: $E\{U_i^{(j)}U_{i'}^{(j)}\}=n_j(n_j-1)/[N(N-1)]$, $(i\neq i')$. The relationship between the samples is defined by:

$$P\{U_i^{(j)}=U_{i'}^{(j')}=1\}=E\{U_i^{(j)}U_{i'}^{(j')}\}=\begin{cases} v_{jj'} : i=i' \\ w_{jj'} : i\neq i' \end{cases},$$

where it is assumed that the expectation depends only on whether $i=i'$ and not on the specific values of i and i'. If we define: $n_{jj'}=n_jn_{j'}/(N\,v_{jj'})$, (where $n_{jj}=n_j$ and it is assumed for the time being that $v_{jj'}\neq 0$), then it is easy to see that:

$$w_{jj'}=\frac{n_jn_{j'}(n_{jj'}-1)}{n_{jj'}N(N-1)} .$$

The statistic considered will be the sample mean: $\underset{\sim}{\bar{y}}_t'=(\bar{y}_t^{(1)},\bar{y}_t^{(2)},\ldots,\bar{y}_t^{(p)})$, where:

$\bar{y}_t^{(j)}=\Sigma_i U_i^{(j)}Y_{it}^{(j)}/n_j$, as an estimate of the population mean: $\underset{\sim}{\bar{Y}}'=(1/N)\sum_{i=1}^{N}\underset{\sim}{Y}_i'$, where: $\underset{\sim}{Y}_i'=E_t\{\underset{\sim}{Y}_{it}'\}$ is the expected response for the i-th element over all trials. It will be assumed that there is no response bias, so that: $Y_i^{(j)}=E_t\{Y_{it}^{(j)}\}=E_t\{Y_{it}^{(j)}|U_i^{(j)}=1\}$. The variance-covariance matrix of $\bar{y}_t$ can be decomposed, as usual, as follows:

$$\underset{\sim}{V}=E_t\{(\underset{\sim}{\bar{y}}_t-\underset{\sim}{\bar{Y}})(\underset{\sim}{\bar{y}}_t-\underset{\sim}{\bar{Y}})'\}=\underset{\sim}{RV}+\underset{\sim}{SV}+2\underset{\sim}{\overline{IV}} . \qquad (1)$$

Letting $\underset{\sim}{\bar{y}}'=(\bar{y}^{(1)},\bar{y}^{(2)},\ldots,\bar{y}^{(p)})$ and $\bar{y}^{(j)}=\sum_{i=1}^{N}U_i^{(j)}Y_i^{(j)}/n_j$ be the sample mean of the expected values for the j-th component, this decomposition reflects the three sources of variation as follows:

Response Variance: $\underset{\sim}{RV}=E_t\{(\underset{\sim}{\bar{y}}_t-\underset{\sim}{\bar{Y}})(\underset{\sim}{\bar{y}}_t-\underset{\sim}{\bar{Y}})'\}$

Sampling Variance: $\underset{\sim}{SV}=E_t\{(\underset{\sim}{\bar{y}}-\underset{\sim}{\bar{Y}})(\underset{\sim}{\bar{y}}-\underset{\sim}{\bar{Y}})'\}$

Interaction: $2\underset{\sim}{IV}=\underset{\sim}{IV}+\underset{\sim}{IV}'$, where: $\underset{\sim}{IV}=E_t\{(\underset{\sim}{\bar{y}}_t-\underset{\sim}{\bar{Y}})(\underset{\sim}{\bar{y}}-\underset{\sim}{\bar{Y}})'\}$.

In the following, $V^{(j,j')}$, the (j,j') element of $\underset{\sim}{V}$, will be expressed by further decomposing each of its components:

$$V^{(j,j')}=RV^{(j,j')}+SV^{(j,j')}+2\,\overline{IV}^{(j,j')} . \qquad (2)$$

Define the simple response variance as:

$$SRV_i^{(jj')}=E_t\{(Y_{it}^{(j)}-Y_i^{(j)})(Y_{it}^{(j')}-Y_i^{(j')})|U_i^{(j)}=U_i^{(j')}=1\}$$

and:

$$SRV^{(jj')}=(1/N)\Sigma_i\,SRV_i^{(jj')} \qquad (3)$$

and, similarly, the correlated response variance as:

$$CRV_{ii'}^{(jj')}=E_t\{(Y_{it}^{(j)}-Y_i^{(j)})(Y_{i't}^{(j')}-Y_{i'}^{(j')})|U_i^{(j)}U_{i'}^{(j')}=1\}$$

and:

$$CRV^{(jj')}=\frac{1}{N(N-1)}\sum_{i\neq i'}CRV_{ii'}^{(jj')} , \qquad (4)$$

*This research was carried out while the author was Visiting Associate Professor at the University of North Carolina at Chapel Hill and at the Research Triangle Institute and was sponsored by U.S. Bureau of the Census Contract No. 1-35096.

(where the conditional expectations are well defined, since we assumed $v_{jj'}=P\{U_i^{(j)}=U_i^{(j')}=1\} \neq 0$ and it can easily be seen that $w_{jj'}=P\{U_i^{(j)}=U_{i'}^{(j')}=1\} \neq 0$, except in the trivial case $n_j=n_{j'}=1$).

The response variance component can be shown to be:

$$RV^{(jj')}=(1/n_{jj'})[SRV^{(jj')}+(n_{jj'}-1)CRV^{(jj')}]. \quad (5)$$

If we define the simple sampling variance as:

$$SSV^{(jj')}=\frac{1}{N-1}\Sigma_i(Y_i^{(j)}-\bar{Y}^{(j)})(Y_i^{(j')}-\bar{Y}^{(j')}), \quad (6)$$

the sampling variance component can be written as:

$$SV^{(jj')}=(1/n_{jj'})(1-\frac{n_{jj'}}{N})SSV^{(jj')}. \quad (7)$$

The interaction component, $IV^{(jj')}$, which reflects the inter-relationship of sampling and response errors, is non-zero if:

$$Y_{ii'}^{(jj')}=E_t\{Y_{it}^{(j)}|U_i^{(j)}=U_i^{(j')}=1\} \neq Y_i^{(j)}.$$

If we define simple interaction between response and sampling deviations for the same units as:

$$SIV^{(jj')}=(1/N)\Sigma_i(Y_{ii}^{(jj')}-Y_i^{(j)})Y_i^{(j')}$$

and: $\overline{SIV}^{(jj')}=1/2(SIV^{(jj')}+SIV^{(j'j)})$; (8)

and define the simple correlated interaction between response and sampling deviations for different units as:

$$SCIV^{(jj')}=\frac{1}{N(N-1)}\sum_{i\neq i'}(Y_{ii'}^{(jj')}-Y_i^{(j)})Y_{i'}^{(j')}$$

and: $\overline{SCIV}^{(jj')}=1/2(SCIV^{(jj')}+SCIV^{(j'j)})$. (9)

Then: $$\overline{IV}^{(jj')}=(1/n_{jj'})[\overline{SIV}^{(jj')}+(n_{jj'}-1)\overline{SCIV}^{(jj')}]. \quad (10)$$

Substituting (5), (7) and (10) in (2) we obtain:

$$V^{(jj')}=\frac{1}{n_{jj'}}\{SRV^{(jj')}+(n_{jj'}-1)CRV^{(jj')}$$
$$+(1-n_{jj'}/N)SSV^{(jj')}+2[\overline{SIV}^{(jj')}$$
$$+(n_{jj'}-1)\overline{SCIV}^{(jj')}]\} , \quad (11)$$

where the last two terms drop out if $Y_{ii'}^{(jj')}=Y_i^{(j)}$ for all (i,i'). It should be noted that this expression has the same form as that given by Koch [4] for the case where all components are measured on the same sample, with only the sample size n replaced by $n_{jj'}$. Thus, methods proposed by Chai [2] and Bailar and Dalenius [1] to estimate the population parameters on the basis of a single sample can be applied.

For the special case, $v_{jj'}=0$ $(j\neq j')$, i.e. non-overlapping samples for the j-th and j'-th component, the resulting modification of the decomposition, (for $j\neq j'$), is:

$$V^{(jj')}=SCRV^{(jj')}+IRV^{(jj')}-\frac{1}{N}SSV^{(jj')}+2\overline{SCIV}^{(jj')}, \quad (12)$$

which is the limit of (11) as $1/n_{jj'}$ goes to zero. It should be noted that $V^{(jj')}$ is independent of sample size, in this case.

2. Application to Complex Estimators: The above model can easily be applied to a variety of sampling designs with respect to the relationship between samples for different components. In the following, the application of the extended model to complex estimators is considered. For the sake of algebraic simplicity, we shall assume no interaction between sampling and response deviations, in the sense that $Y_{ii'}^{(jj')}=Y_i^{(j)}$ for all $i,i'=1,\ldots,N$ and for all $j,j'=1,\ldots,p$. The results can easily be extended to the case of non-zero interactions. The variance of a linear combination of the sample means: $\bar{y}_{t\ell}=\Sigma_j\ell_j\bar{y}_t^{(j)}$, as an estimate of $\bar{Y}_\ell=\Sigma_j\ell_j\bar{Y}^{(j)}$, will be:

$$\mathrm{var}(\bar{y}_{t\ell})=\Sigma_{j,j'}\ell_j\ell_{j'}V^{(jj')} . \quad (13)$$

Similarly the variance of an analytical function of the sample means $g(\bar{y}_t^{(1)},\ldots,\bar{y}_t^{(p)})$ can be approximated by the appropriate Taylor expansion.

We shall consider two variables, X and Y, with observable values at the t-th trial for the i-th unit X_{it}, Y_{it}. Define $X_i=E_t(X_{it})$; $Y_i=E_t(Y_{it})$ and let $\bar{X}=(1/N)\Sigma_iX_i$, $\bar{Y}=(1/N)\Sigma_iY_i$, $\bar{X}_t=(1/N)\Sigma_iX_{it}$ and $\bar{Y}_t=(1/N)\Sigma_iY_{it}$, be the population means of the expected values and of the values for the t-th trial respectively. Let:

$V_{xy}=(1/N)\Sigma_iE_t(X_{it}-X_i)Y_{it}-Y_i)$; and similarly V_{xx} and V_{yy};

$$C_{xy}=\frac{1}{N(N-1)}\sum_{i\neq i'}(X_{it}-X_i)(Y_{i't}-Y_{i'}) ;$$

and similarly C_{xx} and C_{yy};

$S_{xy}=\frac{1}{N-1}\Sigma_i(X_i-\bar{X})(Y_i-\bar{Y})$; and similarly S_{xx} and S_{yy}.

Consider a single simple random sample without replacement of size n, selected from the population of N elements and defined by the indicator random variables U_i $(i=1,2,\ldots,N)$. Let $\bar{x}_t=(1/n)\Sigma_iU_iX_{it}$; $\bar{y}_t=(1/n)\Sigma_iU_iY_{it}$; $\bar{x}=(1/n)\Sigma_iU_iX_i$; and $\bar{y}=(1/n)\Sigma_iU_iY_i$ be the sample means for the t-th trial and for the expected values, respectively.

In the following Y will be the variable for which an estimate of the population mean of expected values, $\bar{Y}$, is required. Measurements of Y are available only for the sample elements at a given trial, t, so that only y_t is known. X will be an auxiliary variable for which measurements are available for the whole population, in one of the following alternative ways:

(a) X is measured at the t-th trial for the population and for the sample. In this case $\bar{x}_t$ and $\bar{X}_t$ are known but not $\bar{x}$ or $\bar{X}$. This would be the most usual case in practice, giving rise to the difference estimate: $\bar{y}_{Dt1}=\bar{y}_t+k(\bar{x}_t-\bar{X}_t)$; or to the ratio estimate: $\bar{y}_{Rt1}=(\bar{y}_t/\bar{x}_t)\bar{X}_t$.

(b) X is measured at the t-th trial for the sample, so that $\bar{x}_t$ is known, and the population mean of the expected values, $\bar{X}$, is known (or alternatively $\bar{X}_t$ has no response error), but $\bar{x}$, the sample mean of expected values, is not known. This case would be rather unusual in practice but may arise if errorless measurements are available for the whole population (or only for its mean) but there is a practical difficulty in matching the sample back, to obtain the values of X_i for the sample elements. The difference estimate for this case is: $\bar{y}_{Dt2}=\bar{y}_t+k(\bar{x}_t-\bar{X})$; and the ratio estimate is: $\bar{y}_{Rt2}=(\bar{y}_y/\bar{x}_t)\bar{X}$.

(c) The expected values of X are known for the whole population and for the sample (or X is assumed to be measured without response error), so that $\bar{x}$ and $\bar{X}$ are known. In this case the difference estimate is $\bar{y}_{Dt3}=\bar{y}_t+k(\bar{x}-\bar{X})$; and the ratio estimate is: $\bar{y}_{Rt3}=(\bar{y}_t/\bar{x})\bar{X}$.

The variances of all these estimates can be obtained from (11) and (13), by considering the vector variables: $\underset{\sim}{Y}'_{it}=(Y_{it},X_{it},X_{it})$, $\underset{\sim}{Y}'_{it}=(Y_{it},X_{it},X_i)$ or $\underset{\sim}{Y}'_{it}=(Y_{it},X_i,X_i)$, for (a), (b) or (c), respectively, with sample sizes: $n_1=n_2=n$ and $n_3=N$ and values of $n_{jj'}$: $n_{11}=n_{12}=n_{22}=n$ and $n_{13}=n_{23}=n_{33}=N$.

The variance of the difference estimates is obtained by applying (13) with $\ell=(1,k,-k)$. The common component of the variances of the three estimates which is independent of k is the variance of the sample mean:

$$n\ \mathrm{var}(\bar{y}_t)=V_{yy}+(n-1)C_{yy}+(1-n/N)S_{yy} \quad . \qquad (14)$$

If we define the intra-trial correlation as:

$$\delta^{(jj')}=\frac{CRV^{(jj')}}{SRV^{(jj')}} \quad (\text{if } SSRV^{(jj')}\neq 0)\ ;$$ so that:

$\delta_{xx}=C_{xx}/V_{xx}$; $\delta_{xy}=C_{xy}/V_{xy}$; $\delta_{yy}=C_{yy}/V_{yy}$, then

minimal variances of the difference estimates are:

$$V_1=\min_k[n\ \mathrm{var}(\bar{y}_{Dt1})]=n\ \mathrm{var}(\bar{y}_t)-(1-f)\frac{[V_{xy}(1-\delta_{xy})+S_{xy}]^2}{V_{xx}(1-\delta_{xx})+S_{xx}} \qquad (15)$$

$$V_2=\min_k[n\ \mathrm{var}(\bar{y}_{Dt2})]=n\ \mathrm{var}(\bar{y}_t)-\frac{[V_{xy}\{1+(n-1)\delta_{xy}\}+(1-f)S_{xy}]^2}{V_{xx}\{1+(n-1)\delta_{xx}\}+(1-f)S_{xx}} \qquad (16)$$

$$V_3=\min_k[n\ \mathrm{var}(\bar{y}_{Dt3})]=n\ \mathrm{var}(\bar{y}_t)-(1-f)S^2_{xy}/S_{xx}\ , \qquad (17)$$

where $f=n/N$.

The last terms of the right hand sides of (15)-(17) represent the possible gains over the sample mean from the use of the three difference estimates. If we assume V_{xy}, S_{xy}, $\delta_{xy}>0$, then for n sufficiently large we will have: $V_2<V_1$ and $V_2<V_3$, so that the greatest gain could be made by using $\bar{y}_{Dt2}$, (i.e. using the trial sample mean, $\bar{x}_t$, and the population expected mean $\bar{X}$), even if errorless (i.e. expected) values are available both for the sample and for the population.

If, however, the response correlations, V_{xy} and C_{xy}, are small, relative to the response variances (V_{xx} and C_{xx}), and n is not too large then $\bar{y}_{Dt3}$ may well have the smallest variance of the three estimates.

The variance of the ratio estimates, $\bar{y}_{Rt}$, will be approximated by applying (13) with $\ell=(1,-R,R)$. Thus (approximately):

$$n\ \mathrm{var}(\bar{y}_{Rt1})\doteq n\ \mathrm{var}(\bar{y}_t)+(1-f)\{R^2[V_{xx}(1-\delta_{xx})+S_{xx}]-2R[V_{xy}(1-\delta_{xy})+S_{xy}]\} \qquad (18)$$

$$n\ \mathrm{var}(\bar{y}_{Rt2})\doteq n\ \mathrm{var}(\bar{y}_t)+R^2\{V_{xx}[1+(n-1)\delta_{xx}]+(1-f)S_{xx}\}-2R\{V_{xy}[1+(n-1)\delta_{xy}]+(1-f)S_{xy}\}\ ; \qquad (19)$$

$$n\ \mathrm{var}(\bar{y}_{Rt3})\doteq n\ \mathrm{var}(\bar{y}_t)+(1-f)[R^2S_{xx}-2RS_{xy}] \quad . \qquad (20)$$

The conditions under which the ratio estimates are better than the sample mean, i.e. $\mathrm{var}(\bar{y}_{Rt})<\mathrm{var}(\bar{y}_t)$, are then:

$$R<2\,\frac{V_{xy}(1-\delta_{xy})+S_{xy}}{V_{xx}(1-\delta_{xx})+S_{xx}} \quad , \text{ for } \bar{y}_{Rt1}\ ;$$

$$R<2\,\frac{V_{xy}[1+(n-1)\delta_{xy}]+(1-f)S_{xy}}{V_{xx}[1+(n-1)\delta_{xx}]+(1-f)S_{xx}} \quad , \text{ for } \bar{y}_{Rt2}\ ;$$

$$R<2\frac{S_{xy}}{S_{xx}} \quad , \text{ for } \bar{y}_{Rt3} \tag{21}$$

If response correlations are small, relative to the correlations between expected values, the conditions (21) for $\bar{y}_{Rt1}$ and $\bar{y}_{Rt2}$ may be more stringent than the condition for $\bar{y}_{Rt3}$ and in particular if:

$$\frac{V_{xy}(1-\delta_{xy})}{V_{xx}(1-\delta_{xx})} < \frac{S_{xy}}{S_{xx}} \quad , \text{ for } \bar{y}_{Rt2} \text{ and if:}$$

$$\frac{V_{xy}[1+(n-1)\delta_{xy}]}{V_{xx}[1+(n-1)\delta_{xx}]} < \frac{S_{xy}}{S_{xx}} \quad , \text{ for } \bar{y}_{Rt1} \quad . \tag{22}$$

The last relationship would always hold if $\delta_{xy}<\delta_{xx}$ for large enough n.

Comparing $\bar{y}_{Rt1}$ and $\bar{y}_{Rt2}$, we have $\mathrm{var}(\bar{y}_{Rt2})>\mathrm{var}(\bar{y}_{Rt1})$ if:

$$R>2\frac{V_{xy}[1+(N-1)\delta_{xy}]}{V_{xx}[1+(N-1)\delta_{xx}]} = 2\frac{E_t\{(\bar{Y}_t-\bar{Y})(\bar{X}_t-\bar{X})\}}{E_t\{(\bar{X}_t-\bar{X})^2\}} \quad . \tag{23}$$

Thus, if the correlation between the trial means, $\bar{X}_t$ and $\bar{Y}_t$, is small, the ratio estimator $\bar{y}_{Rt2}$ would be preferred to $\bar{y}_{Rt1}$, i.e. the trial population, $\bar{X}_t$, should be used rather than the errorless expected population mean, $\bar{X}$, for blowing-up the trial sample ratio, $\bar{y}_t/\bar{x}_t$, even if $\bar{X}$ is known.

3. Application to Sampling on Two Occasions:

Let samples of the same sizes, n, be selected on each of two occasions such that $m(<n)$ elements are matched and measured on both occasions and $u(=n-m)$ are unmatched. Let Y be the variable for the second (current) occasion and X for the first and consider the unbiased estimate from the t-th trial for $\bar{Y}$:

$$\bar{y}'_t=a\bar{y}_{mt}+(1-a)\bar{y}_{ut}+b\bar{x}_{mt}-b\bar{x}_{ut} \quad , \tag{24}$$

where $\bar{y}_{mt}$, $\bar{x}_{mt}$ are the sample means for the matched part, $\bar{y}_{ut}$, $\bar{x}_{ut}$ are the sample means for the unmatched parts and a, b are any constants. Set: $Y'_{it}=(Y_{it},Y_{it},X_{it},X_{it})$ with: $n_1=n_3=m$; $n_2=n_4=u$; $v_{12}=v_{14}=v_{23}=v_{24}=v_{34}=0$; $v_{13}=m/N$; $n_{13}=m$.

If we define

$$T_y^2=V_{yy}-C_{yy}+S_{yy} \; ; \; T_x^2=V_{xx}-C_{xx}+S_{xx} \; ; \text{ and}$$

$$\rho=(V_{xy}-C_{xy}+S_{xy})/(T_xT_y) \; ,$$

then the minimal variance of (24) is:

$$\min_{a,b,U}[\mathrm{var}(\bar{y}'_t)]=C_{yy}-S_{yy}/N+\frac{T_y^2}{n}\cdot\frac{(1+\sqrt{1-\rho^2})}{2} \; , \tag{25}$$

where $U=u/n$.

The variance of the simple sample estimate for the second occasion, $\bar{y}_t=\frac{1}{n}(m\bar{y}_{mt}+u\bar{y}_{ut})$, is:

$$\mathrm{var}(\bar{y}_t)=C_{yy}-S_{yy}/N+T_y^2/n \quad . \tag{26}$$

The factor $\frac{1}{2}(1+\sqrt{1-\rho^2})$ in (25) represents the reduction in the third term of the variance of the simple estimate, (26), obtained by using matched samples and is the same in form as obtained in classical sampling theory without response errors (e.g. in Cochran [3]). However, if the correlated response error C_{yy} is large relative to T_y^2, the matching will not significantly reduce the total variance.

4. References:

[1] Bailar, A. and Dalenius, T. "Estimating the Response Variance Components of the U.S. Bureau of Census' Survey Model." Sankhya B, 31 (1969), 341-60.

[2] Chai, J.J. "Correlated Measurement Errors and the Least Squares Estimator of the Regression Coefficient." Jour. Amer. Statist. Assoc. 66 (1971), 478-83.

[3] Cochran, W.G., Sampling Techniques, 2nd edition, Wiley, 1963, 342-4.

[4] Koch, G.G. "An Alternative Approach to the Multivariate Response Error Models, with Applications to Estimators Involving Subclass Means." (1972). Submitted for publication.

[5] Hansen, M.H., Hurwitz, W.N. and Bershad, M.A. "Measurement Errors in Censuses and Surveys." Bull. Inter. Statist. Inst. 38: 2, (1961), 359-74.