USING PREDICTION-ORIENTED SOFTWARE FOR SURVEY ESTIMATION - PART II: RATIOS OF TOTALS

James R. Knaub, Jr.
US Dept. of Energy, Energy Information Administration, EI-53.1

KEYWORDS:
survey sampling; prediction; estimation; imputation; variance estimation; ratios of totals

ABSTRACT: This article is an extension of Knaub (1999), "Using Prediction-Oriented Software for Survey Estimation," which dealt with the estimation of totals and subtotals and the corresponding estimates of variance in the presence of 'missing data,' whether missing as part of a model-based sampling scheme, or as a result of nonresponse in a census or in any sample survey. The current article deals with ratios of totals. An example from the electric power industry would be the estimation of revenue per kilowatthour and its associated variance estimate. As in Knaub (1999), the goal is to produce such estimates by making use of currently available software in which the model can be quickly and easily modified, and the data may be stored in such a manner that they may be easily recategorized for purposes of publishing various aggregations of the data with corresponding variance estimates.

SOME APPLICATIONS:

A great advantage with this new method is that it is easy to store and manipulate data. Sometimes, published table results differ because "combined" estimates are used in one table or part of a table, and "separate" estimates are used in another. (See Cochran (1977) and Hansen, Hurwitz and Madow (1953).) Also, statistical agencies may present data in different formats, in different tables. It is cumbersome to aggregate data one way to estimate subtotals for one table and another way to estimate subtotals for overlapping areas for another table. The grand totals, for example, would differ. In the case of publishing subtotals for (Bureau of the) Census division regions, consider that Census divisions are groups of States. However, North American Electric Reliability Regions (NERC Regions) have boundaries that cut through States. Further, NERC boundaries recently moved. If imputed values were substituted for each 'missing' observation, using the largest, relatively homogeneous set of data available for each prediction, then they could be aggregated however desired, and using the method of Knaub(1999) and this article, standard errors may be estimated for any aggregation.

For establishment surveys, a very strong reason for using cutoff model-based sampling is that the smallest and most numerous establishments may be unable to supply data on a frequent basis with reasonable accuracy. A lot of imputation may be necessary. Resources are another problem. The method of this article, and Knaub(1999), also applies to imputation for census surveys, and may be used to help publish preliminary subtotals/totals and/or ratios of such numbers more timely. For a design-based sample, this method could be used to predict/impute for the missing members of the sample, and then the aggregate level variances for that part could be added to the variance estimates from the design-based sample. (This technique is used elsewhere. See Lee, Rancourt, and Saerndal(1999).)

NEW METHODOLOGY:

As shown in Knaub (1999), any statistical software package that will provide predicted values, a standard error or variance of the prediction error, and the mean square error (MSE) from the analysis of variance, will suffice for estimating (sub)totals and their variances in the presence of 'missing' data, using the method found in that article. The regression weight must be supplied by means of considerations such as those found in Knaub (1997). For purposes of predicting missing numbers, the population should be categorized into the largest, relatively homogeneous sets of data possible. Imputed numbers are then each individually associated with variance related information that can be regrouped according to whatever aggregations one may wish to publish. The current article goes a step farther and associates pairs of numbers whose ratio is of interest, and then assigns covariance information to the pair for later aggregations. As in Knaub (1999), a given aggregation could contain little or no observed values, yet it may be possible to estimate totals or ratios of totals with some usefulness. Thus 'small area statistics' results may be available.

Here we consider $V_L^*(T^* - T)$, the variance of the error when estimating a total. This is a multiple regression form of V_L in Royall and Cumberland (1981), which contained some more robust variance estimates. However, Knaub (1992), page 879, Figure 1 shows that V_L may do very well, and this multiple regression form of this variance estimator has performed well, as in Knaub (1996) and Knaub (1999).

Now, according to Knaub (1999), using $V_L^*(y_i^* - y_i)$ for the variance of the prediction error (see Maddala (1992)), and noting that $V_L^*(T^* - T) = V_L^*(y_i^* - y_i)$ when there is only one missing value, one finds in Knaub (1999) that in general, we may approximate as follows:

$$V_L^*(T^* - T) =$$

$$\delta(N-n)\sum_r\left\{V_L^*\left(y_i^* - y_i\right) - \frac{\sigma_e^{*2}}{w_i}\right\}$$

$$+\sum_r\frac{\sigma_e^{*2}}{w_i} \quad , \quad \text{where, } 0 < \delta < 1$$

(δ = 0.3 may be a fair general use value; further discussion is found in Knaub (1999).)

$\sum_r$ means to sum over the cases with missing data. (See Royall (1970).) σ_e^{*2} (Knaub (1996)) is the estimated variance of the random factor of the residual, e_0 (Knaub (1993, 1995)), where the error term is $e_i = w_i^{-1/2} e_{o_i}$. w_i is the regression weight, and $(N - n)$ is the number of members of the population that are not in the sample.

(Note: As $(N\text{-}n)$ approaches 1, δ approaches 1. However, δ will generally decrease quickly as $(N\text{-}n)$ becomes a little larger.)

After that, Knaub (1999) discusses adjustments for nonsampling error that would be applicable here also, but will not be repeated here.

EXPOSITION: Organization of Basic Method -

Following is an excerpt from Knaub (1999), page 8:
Picture a typical data file as follows, where "EG" is a category for purposes of performing predictions (an "estimation group"), and "PG" is a category for purposes of publishing subtotals (a "publication group"). Each line represents a record for a given member of the population. A y value is an observed (or "collected") value, and y^* is a predicted value. Let $S1_i^2 = V_L^*(y_i^* - y_i)$, the variance of the prediction error, and $S2_i^2 = \sigma_e^{*2}/w_i$, the mean square error divided by the regression weight, for each case, i.

Example of a partial file:

y_i or y_i^*	$S1_i$	$S2_i$	EG	PG(a)	PG(b)	PG(c)
4359	0	0	1	2	1	3
1289	0	0	2	1	4	4
497	20	17	1	1	3	2
317	13	11	1	2	2	2
223	9	8	2	1	3	2

Here, y^*, $V^*_L(y^*_i - y_i)$, and σ_e^{*2}/w_i are estimated for each missing observation within a given "EG" group, using all data in that group. Then every part of an EG within a given PG is treated as a stratum for estimating the total for that PG group. The variance for each stratum is estimated using the $V^*_L(T^* - T)$ formula above, and the total variance estimate is found by adding the strata variance estimates.

The current problem, however, is to extend this to estimating variance for the estimated ratios of such (sub)totals. Let the estimated ratio be T^*_A / T^*_B, and the variance sought is $V^*_L\left(T^*_A / T^*_B\right)$. In the case of totals, estimates of subtotals and their variances within strata are simply added to obtain an estimate of a total and its variance, respectively. In the case of an estimated ratio of totals, the numerator and denominator are estimated separately, adding stratum components until the estimates of numerator and denominator are completed, and then the estimated ratio is found. For the estimated variance of this estimated ratio, an estimated variance for the numerator, and an estimated variance for the denominator, and a covariance estimate will each have to be constructed from the strata estimates, and then applied to the overall estimations of the ratio and its variance. To do this, however, in a manner that is flexible to changes in PG categorizations, $V^*_L(y^*_i - y_i)$, and σ_e^{*2}/w_i will be needed for each data point associated with the numerator and the denominator, and a fifth number, a covariance component, will be needed, for each related pair of missing data points in the population. This is very little data to have to store and yet leave such flexibility in the publication process for data aggregations.

For one stratum, the estimation of the ratio, T_A / T_B, designated T^*_A / T^*_B, is straightforward. We simply use $T^* = \sum_S y_i + \sum_r y^*_i$ (Royall (1970)) for the numerator, and repeat the application for the denominator. However, variance estimation is more involved and will be discussed below. Further, when considering more than one stratum, an estimated total can be found by adding the subtotal estimates by strata, and similarly for the estimated variance of the total. The estimation of the ratio of totals is also straightforward, but would now involve a more complicated variance formula. However, variance estimation would still rely on only the five stored numbers for every pair of missing data points, plus information designating data categories. (See the table at the end of the next section.)

VARIANCE ESTIMATION:

Starting with the case of a single stratum:

Knaub (1994) is largely a review of and relies heavily upon P.S.R.S. Rao (1992) for covariance formulae associated with the variance of a ratio of variables. Here, however, as in Knaub (1999), the thrust is somewhat different. Here the emphasis is on simplicity of operation, including easily revised models and the association of all information at the individual (pairs of) point(s) level that will be needed to estimate ratios and their variances at any level of aggregation.

As in Knaub (1994), $$\frac{V^*_L\left(T^*_A / T^*_B\right)}{\left(T^*_A / T^*_B\right)^2} =$$

$$\frac{V^*_L\left(T^*_A\right)}{T_A^{*2}} + \frac{V^*_L\left(T^*_B\right)}{T_B^{*2}} - 2\frac{COV^*_L\left(T^*_A, T^*_B\right)}{T^*_A\ T^*_B}$$

Also from Knaub (1994) and Hansen, Hurwitz and Madow (1953), pages 56 to 58,

$$COV^*_L\left(T^*_A, T^*_B\right) = \sum_r COV^*_L\left(y^*_{Ai}, y^*_{Bi}\right) + \cdots$$

which corresponds to

$$V^*_L(T^* - T) > \sum_r V^*_L(y^*_i - y_i)$$ which is explored in Knaub (1999).

By Knaub (1996),

$$\sigma_e^{*2} = \sum_{i=1}^{n} e_{0_i}^2 \,/\, \text{d.f.} = \sum_{i=1}^{n} w_i e_i^2 \,/\, \text{d.f,}$$

where $e_i = y_i - y_i^*$ = residual and d.f. is the number of degrees of freedom, so

$$\text{COV}_{\text{L}}^*(y_{\text{A}i}^*, y_{\text{B}i}^*) = \frac{\sigma^*_{e;y_{\text{A}},y_{\text{B}}}}{w_{\text{A}j}^{0.5} w_{\text{B}j}^{0.5}} + \cdots$$

$$= \frac{\sum^{n} w_{\text{A}j}^{0.5} e_{\text{A}j} w_{\text{B}j}^{0.5} e_{\text{B}j}}{w_{\text{A}j}^{0.5} w_{\text{B}j}^{0.5} (\text{d.f.})} + \cdots$$

and therefore

$$\text{COV}_{\text{L}}^*(\text{T}_{\text{A}}^*, \text{T}_{\text{B}}^*) \approx$$

$$\left[\sum_r \frac{\sigma^*_{e;y_{\text{A}},y_{\text{B}}}}{w_{\text{A}i}^{0.5} w_{\text{B}i}^{0.5}}\right] \left[\frac{\text{V}_{\text{L}}^*(\text{T}_{\text{A}}^*)}{\sum_r \frac{\sigma_{\text{A}e}^{*2}}{w_{\text{A}i}}} \cdot \frac{\text{V}_{\text{L}}^*(\text{T}_{\text{B}}^*)}{\sum_r \frac{\sigma_{\text{B}e}^{*2}}{w_{\text{B}i}}}\right]^{1/2}$$

where $\left[\sum_r \frac{\sigma^*_{e;y_{\text{A}},y_{\text{B}}}}{w_{\text{A}i}^{0.5} w_{\text{B}i}^{0.5}}\right]$ =

$$\left[\sum_r w_{\text{A}j}^{-0.5} w_{\text{B}j}^{-0.5}\right] \left[\sum^{n} w_{\text{A}j}^{0.5} e_{\text{A}j} w_{\text{B}j}^{0.5} e_{\text{B}j} / \text{d.f.}\right]$$

(Note: for $e_{\text{A}i}$, and $e_{\text{B}i}$, one can save $y_i - y_i^*$ in each case (A and B) in another file.)

So, in addition to producing $\text{V}_{\text{L}}^*(y_i^* - y_i)$, and σ_e^{*2} / w_i , for each 'missing' number, also save $\frac{\sigma^*_{e;y_{\text{A}},y_{\text{B}}}}{w_{\text{A}i}^{0.5} w_{\text{B}i}^{0.5}}$ for each pair of corresponding, missing numbers. So, using $\frac{\sigma^*_{e;y_{\text{A}},y_{\text{B}}}}{w_{\text{A}i}^{0.5} w_{\text{B}i}^{0.5}} = S3_i^2$, an example of the requisite data file follows.

Note: If your software calculates mean square error, as shown in the code on page 34 in Knaub(1999), this is not adequate when estimating covariance. Each residual needs to be identified.

Example of a partial file:

$y_{\text{A}i}$ or $y_{\text{A}i}^*$	$S1_{\text{A}i}$	$S2_{\text{A}i}$	$y_{\text{B}i}$ or $y_{\text{B}i}^*$	$S1_{\text{B}i}$	$S2_{\text{B}i}$	$S3_i$	EG	PG(a)	PG(b)
4359	0	0	320	0	0	0	1	2	1
1289	0	0	85	0	0	0	2	1	4
497	20	17	35	9	8	3	1	1	3
317	13	11	22	6	5	2	1	2	2
278	10	9	17	3	3	1	1	1	3
223	9	8	14	3	2	1	2	1	3

An example application is found in Knaub(2000).

REFERENCES:

Cochran, W.G. (1977), Sampling Techniques, 3rd ed., John Wiley & Sons.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953), Sample Survey Methods and Theory, Volume II: Theory, John Wiley & Sons.

Knaub, J.R., Jr. (1992), "More Model Sampling and Analyses Applied to Electric Power Data," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 876-881.

Knaub, J.R., Jr. (1993), "Alternative to the Iterated Reweighted Least Squares Method: Apparent Heteroscedasticity and Linear Regression Model Sampling," Proceedings of the International Conference on Establishment Surveys, American Statistical Association, pp. 520-525.

Knaub, J.R., Jr. (1994), "Relative Standard Error for a Ratio of Variables at an Aggregate Level Under Model Sampling," Proceedings of the Section on Survey Research Methods, Vol. I, American Statistical Association, pp. 310- 312.

Knaub, J.R., Jr. (1995), "A New Look at 'Portability' for Survey Model Sampling and Imputation," Proceedings of the Section on Survey Research Methods, Vol. II, American Statistical Association, pp. 701-705.

Knaub, J.R., Jr. (1996), "Weighted Multiple Regression Estimation for Survey Model Sampling," InterStat, May 1996, http://interstat.stat.vt.edu. (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 1996.)

Knaub, J.R., Jr. (1997), "Weighting in Regression for Use in Survey Methodology," InterStat, April 1997, http://interstat.stat.vt.edu. (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 1997.)

Knaub, J.R., Jr. (1999), "Using Prediction-Oriented Software for Survey Estimation," InterStat, August 1999, http://interstat.stat.vt.edu, partially covered in "Using Prediction-Oriented Software for Model-Based and Small Area Estimation," in ASA Survey Research Methods Section proceedings, 1999.

Knaub, J.R., Jr. (2000), "Using Prediction-Oriented Software for Survey Estimation - Part II: Ratios of Totals," InterStat, June 2000, http://interstat.stat.vt.edu. Article is a longer version of this paper.

Lee, H., Rancourt, E., and Saerndal, C.-E. (1999), "Variance Estimation from Survey Data Under Single Value Imputation," presented at the International Conference on Survey Nonresponse, Oct. 1999, to be published in a monograph.

Maddala, G.S. (1992), Introduction to Econometrics, 2nd ed., Macmillan Pub. Co.

Rao, Poduri S.R.S. (1992), unpublished letters, Aug. - Oct. 1992, on covariances associated with three Royall and Cumberland model sampling variance estimators.

Royall, R.M. (1970), "On Finite Population Sampling Theory Under Certain Linear Regression Models," Biometrika, 57, pp. 377-387.

Royall, R.M. and Cumberland, W.G. (1981), "An Empirical Study of the Ratio Estimator and Estimators of its Variance," Journal of the American Statistical Association, 76, pp. 66-88.

Variance formulation summary:

Per Knaub (1994) and Hansen, Hurwitz and Madow (1953), sum over $\mathrm{COV}^*_{\mathrm{L}_k}\left(\mathrm{T}^*_{\mathrm{A}_k}, \mathrm{T}^*_{\mathrm{B}_k}\right)$ just as is done for $\mathrm{V}^*_{\mathrm{L}_k}(\mathrm{T}^*_k - \mathrm{T}_k)$ for the case of multiple strata, k.

Thus, we have the following

$$\mathrm{V}^*_{\mathrm{L}}\left(\mathrm{T}^*_{\mathrm{A}}/\mathrm{T}^*_{\mathrm{B}}\right) = \frac{\mathrm{V}^*_{\mathrm{L}}\left(\mathrm{T}^*_{\mathrm{A}}\right)}{\mathrm{T}^{*2}_{\mathrm{B}}} + \frac{\mathrm{T}^{*2}_{\mathrm{A}}\mathrm{V}^*_{\mathrm{L}}\left(\mathrm{T}^*_{\mathrm{B}}\right)}{\mathrm{T}^{*4}_{\mathrm{B}}} - 2\,\frac{\mathrm{T}^*_{\mathrm{A}}\,\mathrm{COV}^*_{\mathrm{L}}\left(\mathrm{T}^*_{\mathrm{A}}, \mathrm{T}^*_{\mathrm{B}}\right)}{\mathrm{T}^{*3}_{\mathrm{B}}} \quad \text{, where}$$

$$\mathrm{T}^*_{\mathrm{A}} = \sum_k \mathrm{T}^*_{\mathrm{A}_k},\quad \mathrm{T}^*_{\mathrm{B}} = \sum_k \mathrm{T}^*_{\mathrm{B}_k},$$

$$\mathrm{V}^*_{\mathrm{L}}\left(\mathrm{T}^*_{\mathrm{A}}\right) = \sum_k \mathrm{V}^*_{\mathrm{L}_k}(\mathrm{T}^*_{\mathrm{A}_k} - \mathrm{T}_{\mathrm{A}_k}),\quad \mathrm{V}^*_{\mathrm{L}}\left(\mathrm{T}^*_{\mathrm{B}}\right) = \sum_k \mathrm{V}^*_{\mathrm{L}_k}(\mathrm{T}^*_{\mathrm{B}_k} - \mathrm{T}_{\mathrm{B}_k}) \quad \text{and}$$

$$\mathrm{COV}^*_{\mathrm{L}}\left(\mathrm{T}^*_{\mathrm{A}}, \mathrm{T}^*_{\mathrm{B}}\right) = \sum_k \mathrm{COV}^*_{\mathrm{L}_k}\left(\mathrm{T}^*_{\mathrm{A}_k}, \mathrm{T}^*_{\mathrm{B}_k}\right), \text{ and remembering that}$$

$$\mathrm{T}^*_{\mathrm{A}} = \sum\nolimits_S y_{\mathrm{A}_i} + \sum\nolimits_r y^*_{\mathrm{A}_i},\quad \mathrm{T}^*_{\mathrm{B}} = \sum\nolimits_S y_{\mathrm{B}_i} + \sum\nolimits_r y^*_{\mathrm{B}_i},$$

$$\mathrm{V}^*_{\mathrm{L}_k}(\mathrm{T}^*_{\mathrm{A}_k} - \mathrm{T}_{\mathrm{A}_k}) = \delta_{\mathrm{A}}\left(N_{\mathrm{A}} - n_{\mathrm{A}}\right)\sum\nolimits_r\left\{\mathrm{V}^*_{\mathrm{L}_k}\left(y^*_{\mathrm{A}_{k_i}} - y_{\mathrm{A}_{k_i}}\right) - \frac{\sigma^{*2}_{\mathrm{A}e}}{w_{\mathrm{A}_i}}\right\} + \sum\nolimits_r \frac{\sigma^{*2}_{\mathrm{A}e}}{w_{\mathrm{A}_i}},$$

$$\mathrm{V}^*_{\mathrm{L}_k}(\mathrm{T}^*_{\mathrm{B}_k} - \mathrm{T}_{\mathrm{B}_k}) = \delta_{\mathrm{B}}\left(N_{\mathrm{B}} - n_{\mathrm{B}}\right)\sum\nolimits_r\left\{\mathrm{V}^*_{\mathrm{L}_k}\left(y^*_{\mathrm{B}k_i} - y_{\mathrm{B}k_i}\right) - \frac{\sigma^{*2}_{\mathrm{B}e}}{w_{\mathrm{B}_i}}\right\} + \sum\nolimits_r \frac{\sigma^{*2}_{\mathrm{B}e}}{w_{\mathrm{B}_i}},$$

and

$$\mathrm{COV}^*_{\mathrm{L}}(\mathrm{T}^*_{\mathrm{A}}, \mathrm{T}^*_{\mathrm{B}}) \approx \left[\sum\nolimits_r w^{-0.5}_{\mathrm{A}j} w^{-0.5}_{\mathrm{B}j}\right]\left[\sum^{n} w^{0.5}_{\mathrm{A}j} e_{\mathrm{A}j} w^{0.5}_{\mathrm{B}j} e_{\mathrm{B}j} / \text{d.f.}\right]\left[\frac{\mathrm{V}^*_{\mathrm{L}}\,(\mathrm{T}^*_{\mathrm{A}})}{\sum_r \frac{\sigma^{*2}_{\mathrm{A}e}}{w_{\mathrm{A}i}}} \cdot \frac{\mathrm{V}^*_{\mathrm{L}}\,(\mathrm{T}^*_{\mathrm{B}})}{\sum_r \frac{\sigma^{*2}_{\mathrm{B}e}}{w_{\mathrm{B}i}}}\right]^{1/2} .$$