Using Prediction-Oriented Software for Survey Estimation - Part III: Full-Scale Study of Variance and Bias

James R. Knaub, Jr.
US Dept. of Energy, Energy Information Administration, EI-53.1, Washington, DC 20585

Key Words: small area estimation, imputation, standard error, bias, inference, graphical editing, regression, models, cutoff sampling

Abstract:
Applications for this method include small area estimation and imputation, with estimates of standard errors of totals. This has been tested and developed, and now enters full-scale testing and implementation. Emphasis will be on the integration of this method into overall survey processing. Advantages include ease in revising models, flexible organization, storage and usage of data, and the ability to maximize the effectiveness of collected data. It integrates well with graphical editing. For purposes of estimation, collected data may be grouped such that each data set contains as many members as may be well defined under a single model per group. That is, each category (group) should be as large as it can be and remain basically homogeneous. Regressor data on the universe are required.

After the models are exercised, there will be either an observed response or an 'imputed' value for each member of the population, which can be rearranged and published, with estimated standard errors, for any subtotals desired. As a matter of practical importance, data tables containing observed and imputed values, illustrated in Knaub(1999) on pages 8, 9 and 22, are very helpful to people processing such data for publication, especially when those processing the data may not be inclined to do statistical analyses. Errors in publishing (sub)totals, caused by duplicate records or 'dropped' records are easier to discover when a data manager can see a table for all members of the universe, which contains either an observed or an imputed number in each case. (One must, however, guard against a customer confusing an imputed number for a reported number for a given establishment.) Scatter graphs used for graphical editing can be used in conjunction with these tables.

Under full-scale testing, more results have become available for a better study of variance estimation, and bias is also studied with instructive results. Other areas illustrated are the appropriateness of using this technique under an extreme condition, and the application of this method across strata.

Introduction:

This work is a continuation of Knaub(1999) and Knaub(2000), where it is shown that any software that performs regression and will allow system calculated values such as residuals to be used in new calculations, can be used to estimate any subtotals and their standard errors. Data may be grouped optimally for estimation purposes, and regrouped for publication purposes. There will be either an observed or an imputed value in each case. Imputed values are associated with two other numbers. The first is the standard error of the prediction error for that imputed number, and the second is the root mean square error divided by the square root of the regression weight. This latter number is needed when estimating standard errors for (sub)totals to be published using this flexible system. The advantage lies in the simplicity with which data may be stored and rearranged for publishing under a variety of categories. This method may be used for inference with model-based sampling, or as an imputation tool for any kind of sample or census survey, for which regressor data are available. For a design-based sample, imputations may be made first and then variance due to imputation added to variance due to the design. (See Lee, Rancourt and Saerndal (2002).)

The regression model may have any number of regressors, with regression weights defined as a function of a single regressor, or combination of regressors. For example:

$y_i = \beta_1 x_{1_i} + \beta_2 x_{2_i} + e_{0_i} \left(0.5x_{1_i} + 1.3x_{2_i}\right)^{\gamma}$, where $e_i = w_i^{-1/2} e_{o_i}$ and w_i is the regression weight such that $w_i^{-1/2} = \left(0.5x_{1_i} + 1.3x_{2_i}\right)^{\gamma}$. In this example, it might be the case that $\hat{y}_i = 0.5x_{1_i} + 1.3x_{2_i}$ is a preliminary estimate of y. The format, $x^{\gamma} = w^{-1/2}$, is well established as quite useful. (See Cochran(1953) and Knaub(1995).) Estimation of the value of gamma indicated by the data is also discussed in Knaub(1993), and Knaub(1997). However, in addition to the best gamma value as indicated by the precise data used in the model at a given time, there are other considerations. These considerations are discussed in Knaub(1999). Also see Knaub(1997). The range of useful gamma values will also be discussed in an upcoming book by K.R.W. Brewer (Brewer(2002)).

It is shown in Knaub(1999) that a good estimate of the variance of the estimate of a subtotal for any stratum is

$$\mathrm{V}_{\mathrm{L}}^{*}(\mathrm{T}^{*}-\mathrm{T})=\delta\,(N-n)\sum_{r}\left\{\mathrm{V}_{\mathrm{L}}^{*}\left(y_i^{*}-y_i\right)-\frac{\sigma_e^{*2}}{w_i}\right\}+\sum_{r}\frac{\sigma_e^{*2}}{w_i},$$

where $0<\delta<1$, and "*r*" indicates summation over the N-n nonrespondents within the stratum. $\mathrm{V}_{\mathrm{L}}^{*}\left(y_i^{*}-y_i\right)$ is the variance of the prediction error (see Maddala(1992)), $\sigma_e^{*2}\big/w_i$ is the MSE divided by the weight, and σ_e^{*2} is the estimated variance of the random factor of the residual, e_0 (see Knaub (1993, 1995)), where the error term is $e_i=w_i^{-1/2}e_{o_i}$. Also, w_i is the regression weight. For highly skewed electric power data, $\delta=0.3$ works very well. (See Knaub(1999) and Knaub(2000).) Further, results have appeared to be more sensitive to γ than to δ.

Each stratum consists of a subsection of the category to be published (publication group, "PG") that also belongs to a part of the population for which a single model was used (an estimation group or "EG"). Thus, the strata are each intersections of the PG with an EG. Variance for the PG is estimated by the total of the estimated variances for each of the strata within the PG.

An important feature of this methodology is the flexibility that it gives to data storage and reconstitution under various categories for subtotaling the results. The estimation groups, EGs, do not need to correspond to the publication groups, PGs. Thus, estimations (imputations) for missing values can be made using optimally grouped EGs for that purpose, regardless as to what PGs are to be shown in data reports. This is also useful with regard to small area estimation, allowing estimation within some strata of a PG that otherwise may not have been possible/practical. The lack of data would result in a large variance for such strata, and perhaps substantial bias, but accuracy would be improved over what would otherwise be obtained, as found in practice at the Energy Information Administration.

Model-bias, particularly for cutoff samples, was a study topic suggested by the American Statistical Association's Committee on Energy Statistics after reviewing this method at a meeting in the Fall of 2000. Cochran(1977) and Hansen, Hurwitz and Madow(1953) discuss the bias in model-assisted design-based estimation of totals, which are shown to diminish with increased sample size, and be proportional to the standard error. For the model-based case, Brewer(2002) discusses conditions under which the bias would be negative. Further, Valliant, Dorfman and Royall(2000) discuss work by Royall and Herson(1973) that uses a polynomial format to generalize a regression model (with one regressor) to show that model-bias can be eliminated by making the sample mean for the regressor, *x*, equal to the population mean of X, when using a model-based sample. This is called a 'balanced sample.' However, this may not always be a practical solution, especially for highly skewed establishment surveys. There are cases where an agency has only wanted to report on the largest entities and ignore the others. Fairness in information disclosure may enter into consideration. Perhaps an establishment might file a complaint about reporting certain data if other establishments of the same 'size' are not required to report as often. Also, trying to implement a balanced sample could introduce more respondent burden than may be allowed. Perhaps more of a problem, if randomization were used, that could result in the need for substantial imputation anyway. Substantial nonsampling error often occurs when attempting to collect timely data from smaller establishments. Cutoff samples have therefore been used for electric power surveys at the Energy Information Administration, and a study of bias when using the new methodology of Knaub(1999) is in order.

Discussion of model-introduced bias in Valliant, Dorfman and Royall(2000) is quite clear: Since a polynomial can be used to fit other distributional forms, the model used can be thought of as one where most terms are not present. Use of such a model is not bad practice in that one should not overspecify, given lack of perfect knowledge. Still, bias may be reduced for the more insightful models. The presence of multiple regressors is a further complication, but the principle is the same. The number of regressors may be varied, as well as γ , and here, δ may be varied. In this study, the focus is upon γ . Compared to γ , δ had little influence on resulting variance estimation. Further, Brewer(2002) indicates a zero intercept is probably best, and this author's research seems to indicate likewise. Also, $x^{\gamma}=w^{-1/2}$ is a general and useful format, as mentioned above. Therefore, for this study of utility generation estimation, γ is a strong influence on the appropriateness of the model, and therefore, a strong influence on model-introduced bias. This should be true in general. The best value to be used for γ , however, can be somewhat elusive. (See Knaub(1997), Knaub(1993) and Knaub(1995).) Is further adjustment for bias possible, or even advisable? That is a subject that is taken up in the case study below.

Extreme Circumstance:
Suppose that a single regressor, or function of regressors

used for y, is z. Suppose further, that this was a mistake: there is no correlation between y and z! A balanced sample would guard against bias under this situation as well, but a cutoff model-based sample should underestimate totals (under-predicting for each missing value) when regression is forced through the origin. This is illustrated in Knaub (2001).

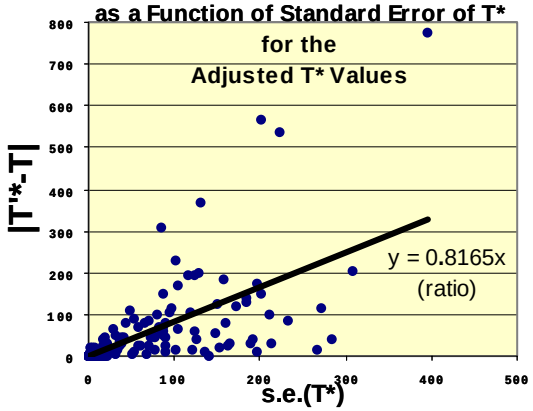
Balanced sample:
Under such an extreme condition, such as mentioned above, a "balanced sample" (Valliant, Dorfman and Royall(2000)), would be useful. However, using a balanced sample may often be better in theory than in practice. Sometimes, in establishment surveys, data customers are only tracking the largest few entities. That may be all that is collected and published. Some published "totals" in official statistics are only the sum of observations in a 'sample,' and it may be difficult to convince the responsible agency that this is not adequate. What is excluded may be small at a high aggregate level (perhaps a national level), but large for some published less aggregate levels (say, State level numbers). Thus estimation for the remaining (many) relatively 'small' establishments may also be important, and therefore a cutoff sample, rather than a truncated universe, may be desired. Using a balanced sample would force the collection of a larger sample size. Like a design-based sample, there would be a number of smaller observations required, which may have to be imputed anyway due to large nonsampling error, or nonresponse.

Case Study (154 'Samples'):
In spite of the possibility of a negative bias shown above, and that found in Brewer(2002), the 'obtained' bias in the following study was positive. Electricity generation data were obtained from utilities, by State, by energy source (for hydroelectric, coal-fired, gas-fired and petroleum-fired generation). There were 154 such categories in this experiment. Data were obtained from a census, with regressor data taken from a previous, similar census and yet another census survey. A standard testing procedure is to simulate a sample by using part of the formerly mentioned census, and that was done here. Here, T represents the total generation actually observed for a given fuel type and State. T* represents the estimate, formed by summing observed values from a cutoff 'sample,' and imputed values for the 'missing' observations. The standard error of T* is thus $\left\{V_L^*(T^* - T)\right\}^{0.5}$ = s.e.(T*). z-values found in Figure 4 are thus $z = (T^*-T)/s.e.(T^*)$.

Two sets of results are considered. In one set, the gamma value was considered by fuel type, and at least part of the remaining apparent positive bias was subtracted from T* (to form a new estimate of T), so that the resulting z-values appeared to be distributed well with regard to variance, and with a nearly symmetric shape, indicating no substantial remaining bias. (See Figure 3.) In that case, the 'adjusted' z is $z' = (T^*-T- c_f)/s.e.(T^*)$, where c_f was a fraction of the s.e.(T*) value for each fuel type, f. (All variables could be subscripted with an "f" here.) The new estimate of T is then $T'^* = T^*- c_f$. However, this might be considered tantamount to over-specification. (See Knaub(1995) with regard to the somewhat fickle nature of gamma.) In the other set of results below, gamma is set at 0.5 (the 'ratio' estimate) for all but the gas-fired cases, which seem quite different, and experimentation showed that gamma was much better set at 0.8 for those instances. There was no further adjustment. (Note that in later applications to monthly testing, where fuel switching, seasonality and higher nonsampling errors can confuse the situation, the ratio estimate, being more robust, appeared useful for gas-fired generation as well.)

Before showing the graphs of z-values for these two sets of results, graphs are presented that show |T*-T| as a function of s.e.(T*). (Note that in the formerly described, or more 'adjusted' results, T* has subtracted from it a fraction of the standard error, varying by fuel type, and may therefore be designated as T'*.) The Excel generated "trend lines" in the figures below automatically assume OLS, so SAS PROC REG was used to estimate the equations using a ratio, model-based estimate. These graphs show that the standard errors for the latter results (less adjusted, only using gamma equal to 0.5 or 0.8), conservatively cover error estimation so that there is little

Figure 1
Absolute Error of (T'*)
as a Function of Standard Error of T* for the Adjusted T* Values



chance of indicating greater accuracy than has been achieved. That could be quite important in official statistics. Also, remember $T^*(\gamma = 0.5)$ appears robust.

Figure 2

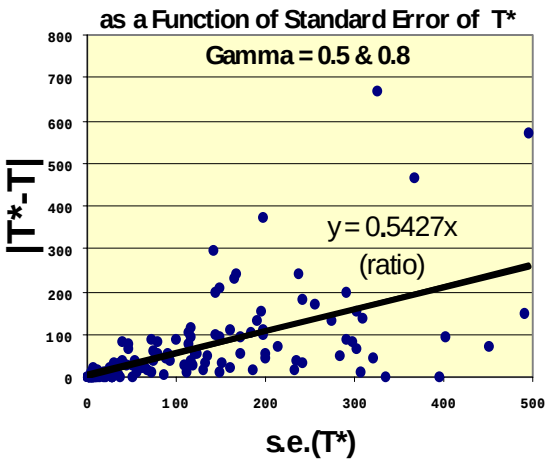


Figure 3

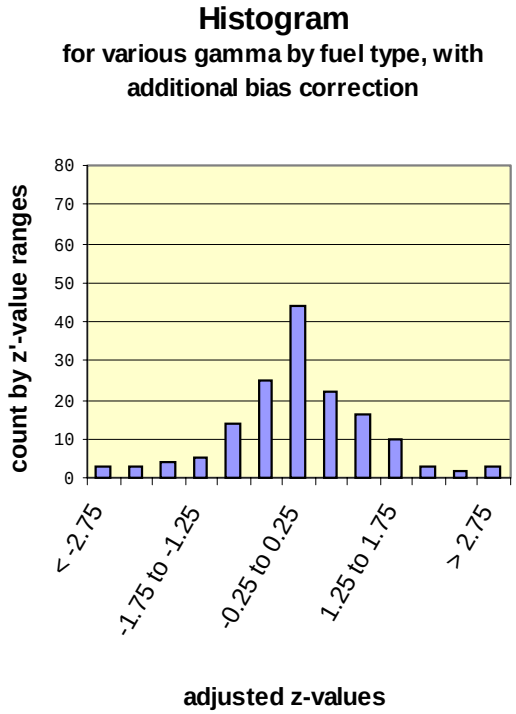
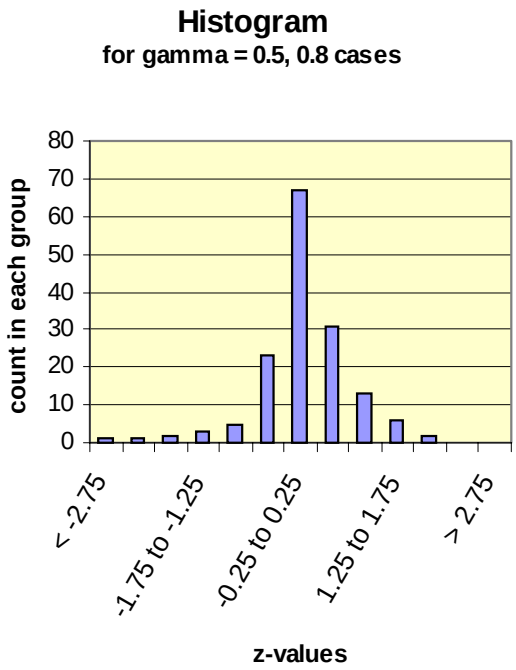


Figure 4



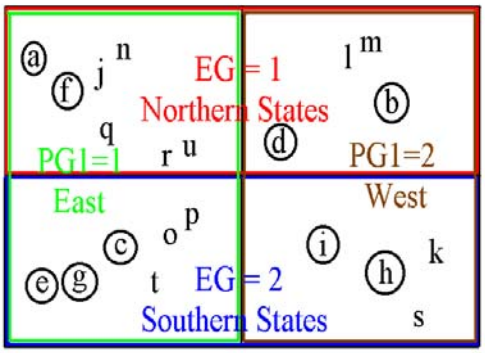
Conclusion with Regard to Case Study:
Use of gamma = 0.5 for all cases except for electricity generated from natural gas, where gamma = 0.8 was used, appears to result in noticeable bias and a slight overestimate of variance. However, as a general indicator of the reliability of the associated estimates of (sub)totals, results appear satisfactory. By avoiding further 'adjustments,' programming may be more generally applicable so that monthly production of reports based on monthly sampled observations may proceed more smoothly. The ratio estimate (gamma=0.5) in particular, has shown robust behavior. (This has also appeared to be the case, in the author's opinion, for other situations encountered over a number of years.) This may contribute toward uninterrupted production of frequently produced data publications.

To put these results into practice, one must consider stratification. Thus the next section describes that procedure.

New Method: Application across strata:
An example of a "partial" data file illustrating the new

method is found on page 8 of Knaub(1999). Figure 5 below represents such data.

Figure 5



- **An 'EG' is an 'estimation group' for purposes of regression modeling.**
- **A 'PG' is a 'publication group' from which a subtotal is to be published.** (There may be any number of such groupings.)
- The data points 'observed' are circled. Those 'imputed' are not circled.
- Estimated totals are easy to obtain. For the estimated total in the western States (PG1 = 2), simply add the observed and imputed numbers for d, l, m, b, *i*, h, k and s.
- For variances, sum the variance estimates from relevant strata.
  (A stratum is the intersection of a 'PG" with an 'EG.')
- Example for a given stratum**:**
  The estimated variance for the southern stratum of PG1 = 2 sums information for k and s, using data from e, g, c, *i* and h in a model applicable to EG = 2. (More data would normally be present, but this is just for purposes of illustration.)

- How to establish the EGs:
  Practical matters, such as considering areas with similar precipitation patterns when collecting data on hydroelectric generation, should be part of this decision making process. 'Trial and error' may be used, comparing model parameters, but caution should be exercised if hypotheses are tested. (See Knaub(1987).) Since p-values are sample size dependent, confidence intervals are often more informative. It is possible that a given PG will contain no collected data at all, yet with wise determination of EGs, a decent estimate for that PG could be obtained. (However, this is not likely, and the estimated variance would normally be large, so it would not be crucial to know the bias. The estimated (sub)total would probably not be publishable.)

Epilogue:
This method is now being implemented for two sample surveys as a small area method, and may be tested as an imputation method for at least one census survey in the near future, possibly to be expanded to several others. Test data results were good, and there is a clear understanding as to the implementation of this method across strata. It can be used for imputation for any kind of survey, including design-based sample surveys. (See Lee, Rancourt and Saerndal(2002).) Applications as a small area technique, as opposed to imputation for a census, may best be accomplished by attention to the gamma values employed, as indicated in Knaub(1999). Use with design-based sampling, and differences between establishment and household surveys and any other possible applications may be addressed by giving attention to the delta value(s) chosen, as well as gamma.

The Energy Information Administration is currently beginning to use this methodology for the ***Electric Power Monthly*** publication, as a means for estimation. Related graphical edits are being implemented on a larger scale, to help identify nonsampling error. Thus the application of models is being expanded.

Further considerations for implementation:
Many practical matters must be taken into account. In some cases, regressor data may not be complete, or some change may have taken place at an establishment which would cause the model to no longer apply to those data. In such a case, data collected from that establishment may be used to represent only that establishment, and should not be used to estimate for 'missing' data. Such responses may be labeled as "ADD-ONs." (Note: For purposes of data editing, it is very useful to graph the data element of interest as a function of a regressor or of a function of regressors. "Add-ons" may not be included in such graphs. Other scatterplots may examine the relationship between data elements when data are not complete for any of these elements but the graphs may still be useful for data editing purposes.) Another practical consideration would be changes in the frame due to company mergers. Making certain that regressor data and current data of interest are matched properly is often far from trivial. That is one more way that an agency with a disorganized approach can find itself in trouble.

Finally, consider nonsampling error. One approach was mentioned in Knaub(1999), on pages 8 and 9. It involves a study that could include noting revisions to observed responses. Another possibility might be a simulation that predicts for successively removed responses and "averages" the predicted standard errors.

Comments on SAS code for implementation:
Code on pages 18 through 23 in Knaub(2001) was extracted from a SAS program written by Dr. Orhan M. Yildiz for application at the Energy Information Administration. Fragments of this code may be useful to various readers. (Also see the shorter, more generalized code on pages 33 and 34 of Knaub(1999).) Note that coefficients and other statistics are determined here by SAS PROC REG, but that other software might be used to perform the same functions.

Acknowledgments:
Thanks to the American Statistical Association's Committee on Energy Statistics for comments that at the least helped lead to a study of bias. Thanks also to Dr. Orhan Yildiz for helpful discussions and also for SAS programming support, and to others at the Energy Information Administration and elsewhere, for helpful discussions.

References:

Brewer, KRW (*forthcoming* 2002), Sampling Basu's elephants: Combining design-based and model-based inference, Arnold: London.

Cochran, W.G. (1953), Sampling Techniques, 1st ed., John Wiley & Sons, (3rd ed., 1977).

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. (1987), "Practical Interpretation of Hypothesis Tests," Vol. 41, No. 3 (August), letter, The American Statistician, pp. 246-247.

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. (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. (1997), "Weighting in Regression for Use in Survey Methodology," InterStat, URL: http://interstat.stat.vt.edu/interstat/intro.html-ssi , April 1997. (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, URL:http://interstat.stat.vt.edu/interstat/intro.html-ssi , August 1999, 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," June 2000, InterStat, URL: http://interstat.stat.vt.edu/interstat/intro.html-ssi . (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 2000.)

Knaub, J.R., Jr. (2001), "Using Prediction-Oriented Software for Survey Estimation - Part III: Full-Scale Study of Variance and Bias," InterStat, URL:http://interstat.stat.vt.edu/interstat/intro.html-ssi , June 2001. (This article is the 'long' version of the current proceedings paper.)

Lee, H., Rancourt, E., and Saerndal, C.-E. (*circa* 2002), "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 in Dec. 2001.

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

Royall, R.M., and Herson, J. (1973), "Robust Estimation in Finite Populations," Journal of the American Statistical Association, 68, pp. 880-889.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000), Finite Population Sampling and Inference, A Predictive Approach, John Wiley & Sons.