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, model-based sampling

Abstract:
Uses for this method include small area estimation and imputation, accompanied by estimates of standard errors. It has been tested and developed, and now enters full-scale testing and implementation. Advantages include ease in revising models, flexible organization, storage and usage of data, and the ability to maximize the effectiveness of collected data. 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 statistically oriented. 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.)

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.

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). The range of useful gamma values will also be discussed in an upcoming book by K.R.W. Brewer (Brewer(2002)).

In general, it is shown in Knaub(1999) that a good estimate of the variance of the estimate of a subtotal for any stratum is

$$V_L^*(T^* - T) = \delta(N-n)\sum_r \left\{ V_L^*\left(y_i^* - y_i\right) - \frac{\sigma_e^{*2}}{w_i} \right\} + \sum_r \frac{\sigma_e^{*2}}{w_i} \quad , \quad \text{where,}$$

$0 < \delta < 1$, and "*r*" indicates summation over the N-n nonrespondents within the stratum. 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, each stratum is an intersection of a 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. 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 .

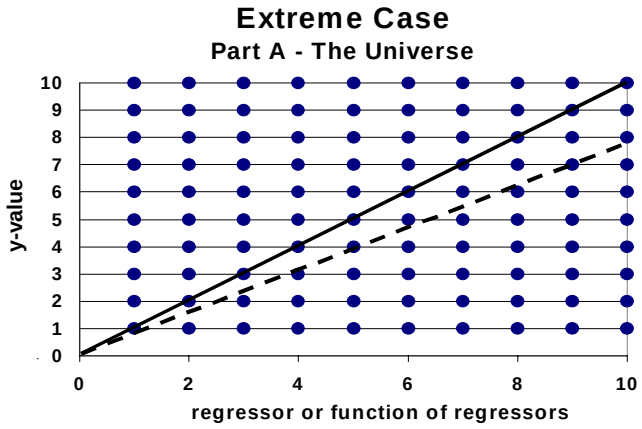
An important feature of this methodology is the flexibility that it gives to data storage and reconstitution under various categories for subtotaling the results. The estimation groups, EGs, do not need to correspond to the publication groups, PGs. Thus, estimations (imputations) for missing values can 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 in general, but particularly with regard to cutoff sampling, was a study topic, the need for which was indicated by the American Statistical Association's Committee on Energy Statistics after reviewing this method. 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 where 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. 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. 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), or any inference from cutoff model-based samples, is in order.

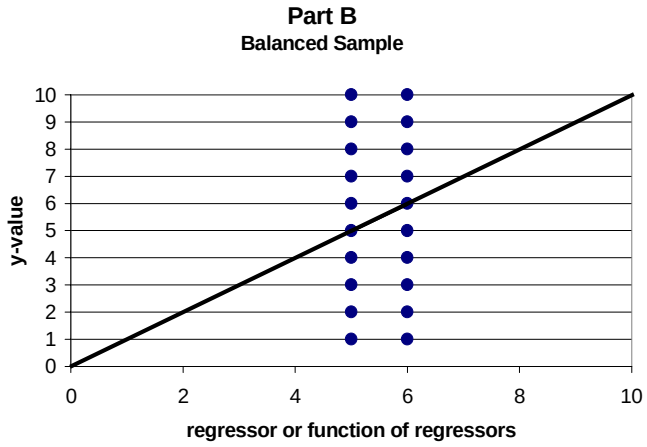
Discussion of model-introduced bias in Valliant, Dorfman and Royall(2000) is quite clear: Since a polynomial can be used to fit other distributional forms, the model used can be thought of as one where most terms are not present. Use of such a model is not bad practice in that one should not overspecify, 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. However, compared to γ, δ had little influence on resulting variance estimation. Thus, γ is the subject of this study. 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 value used for γ, however, can be somewhat volatile. (See Knaub(1997) and Knaub(1995).)

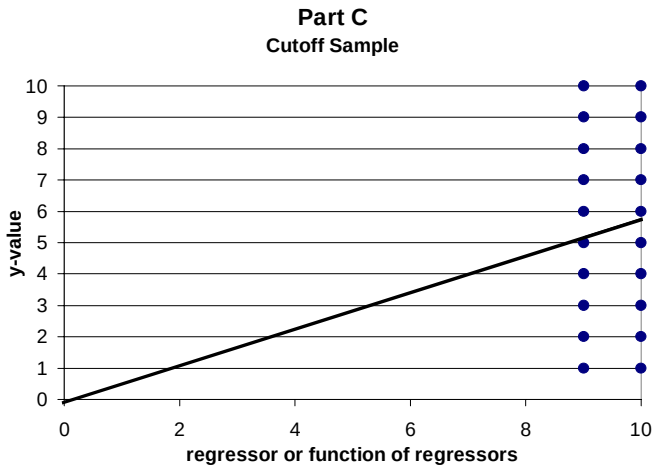
Is further adjustment for bias possible, or even advisable? That is a subject that is taken up in the case study below.

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). This is illustrated below:



The solid line is for $\gamma = 0.5$, the 'ratio' estimate, and the dashed line is for $\gamma = 0$, or OLS regression.





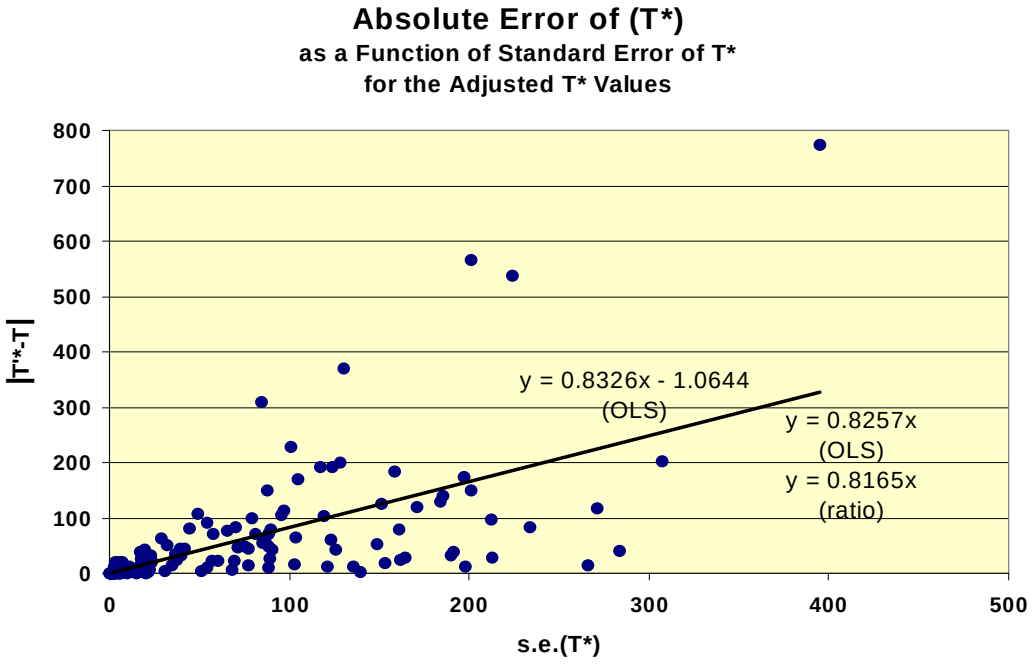
Balanced sample:
Under such an extreme condition, such as shown above, a "balanced sample," Valliant, Dorfman and Royall(2000), would be useful. However, using a balanced sample may often be better in theory than in practice. Sometimes, in establishment surveys, data customers are only tracking the largest 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. What is left out 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 thus a cutoff sample, rather than a truncated universe, may be desirable. 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 nonsampling error.

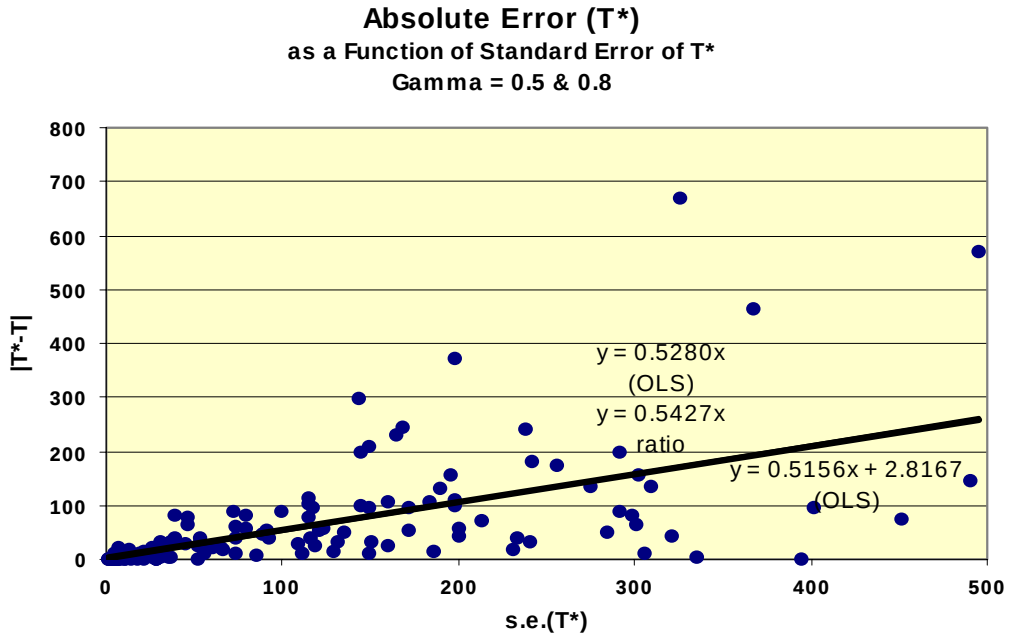
Case Study (154 'Samples'):
In spite of the possibility of a negative bias shown above, and that found in Brewer(2002), the 'obtained' bias in the following study was positive. Electricity generation data were obtained from utilities, by State, by energy source (for hydroelectric, coal-fired, gas-fired 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 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. Simple cutoff samples were formed. 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 the 'sample' and imputed values for the 'missing' observations (below a cutoff). The standard error of T* is thus $\left\{V_L^*(T^* - T)\right\}^{0.5}$ = s.e.(T*). z-values found in the graphs 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 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 reasonably with regard to variance, and with a symmetric shape, indicating no substantial remaining bias. (See the graphs below.) In that case, $z_f' = (T_f^* - T_f - c_f)/s.e.(T_f^*)$, where c_f was a fraction of the $s.e.(T_f^*)$ value for each fuel type, f. For example, z' = (T*- 0.3se(T*) - T)/se(T*) , which indicates that T'* = T*- 0.3se(T*). 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 better set at 0.8 for those instances. There was no further adjustment.

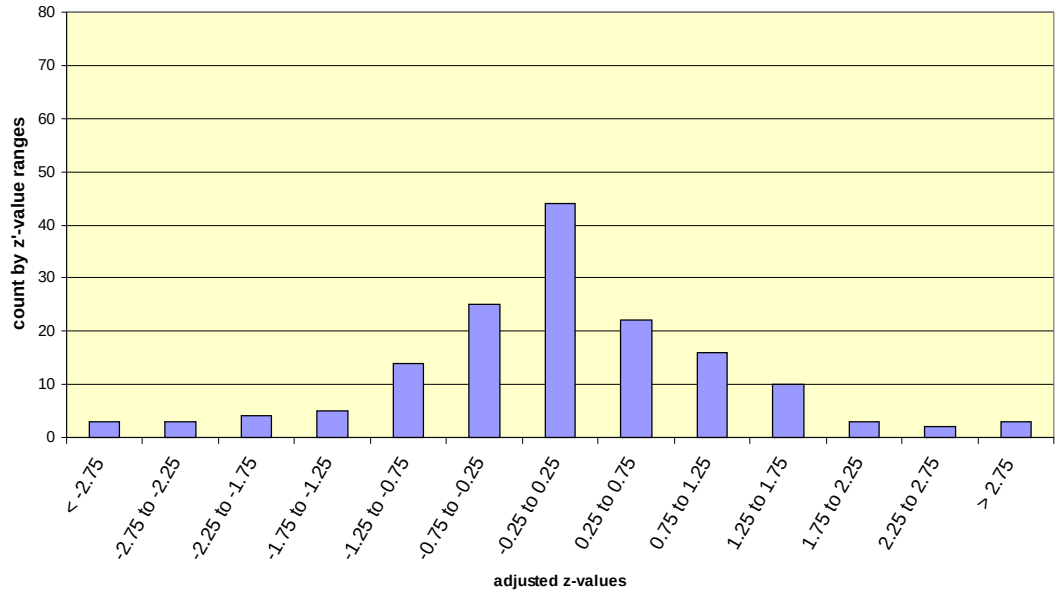
Before showing the graphs of z-values for these two sets of results, graphs are presented that show |T'*-T| or |T*-T| as a function of s.e.(T*). (Note that in the formerly described, or more 'adjusted' results, T* has subtracted from it a fraction of the standard error, varying by fuel type. Even without this, T* and se(T*) are different in the two cases because the γ values are different.) The Excel

generated “trend lines” automatically assume OLS, so SAS PROC REG was used to estimate the slopes of these lines using a ratio, model-based estimate too. These graphs show that the standard errors for the latter results (the more general case, only using gamma equal to 0.5 or 0.8), are generally overestimated, so that there is less chance of indicating greater accuracy than has been achieved. That is quite important in official statistics. Also gamma = 0.5 appears to provide robust estimates of totals, based on substantial experience.

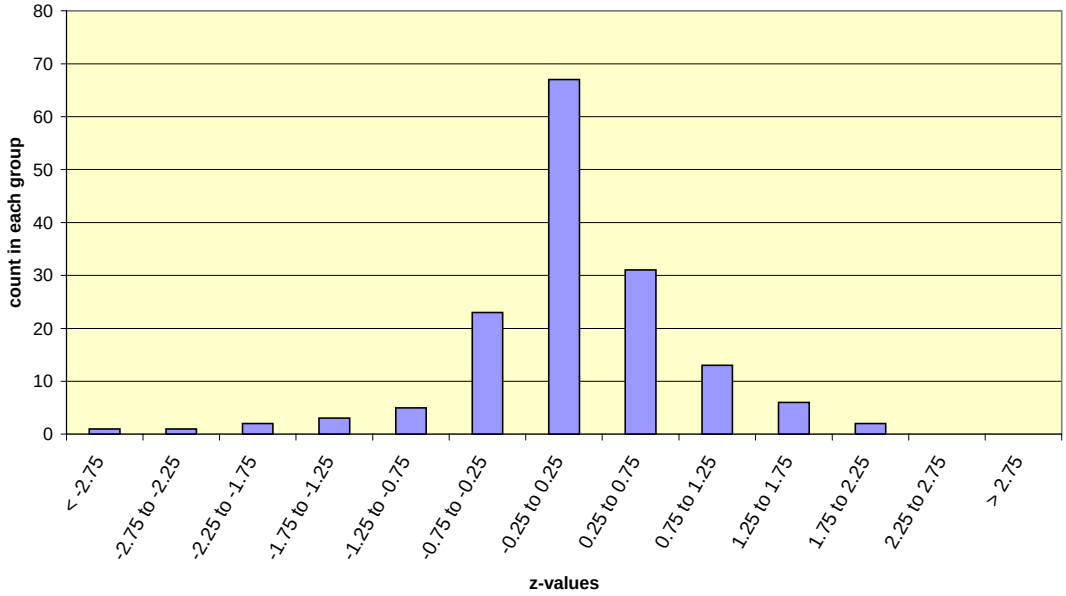


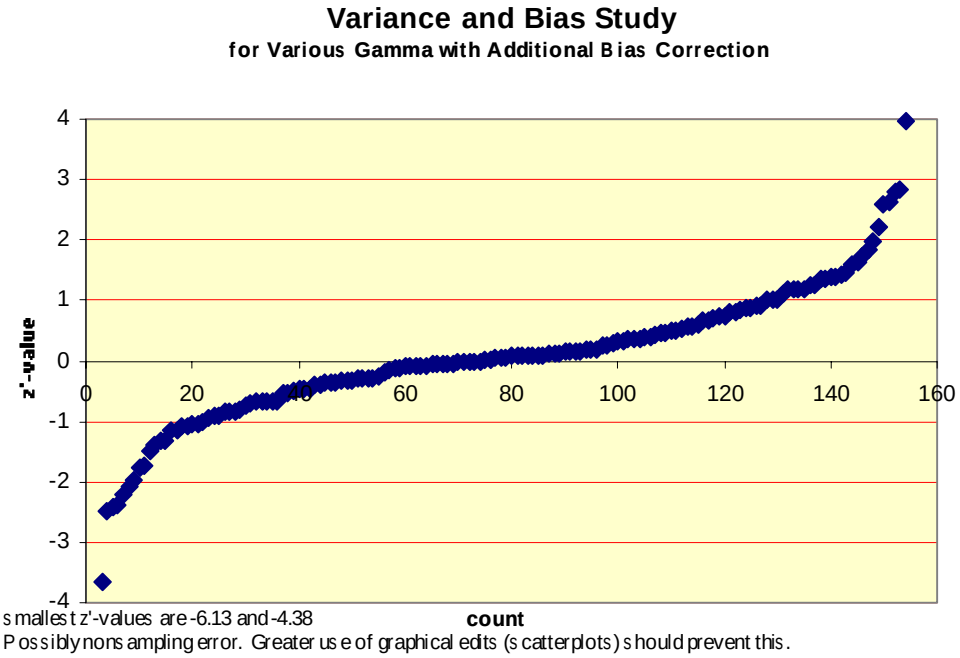


Histogram
for various gamma by fuel type, with additional bias correction



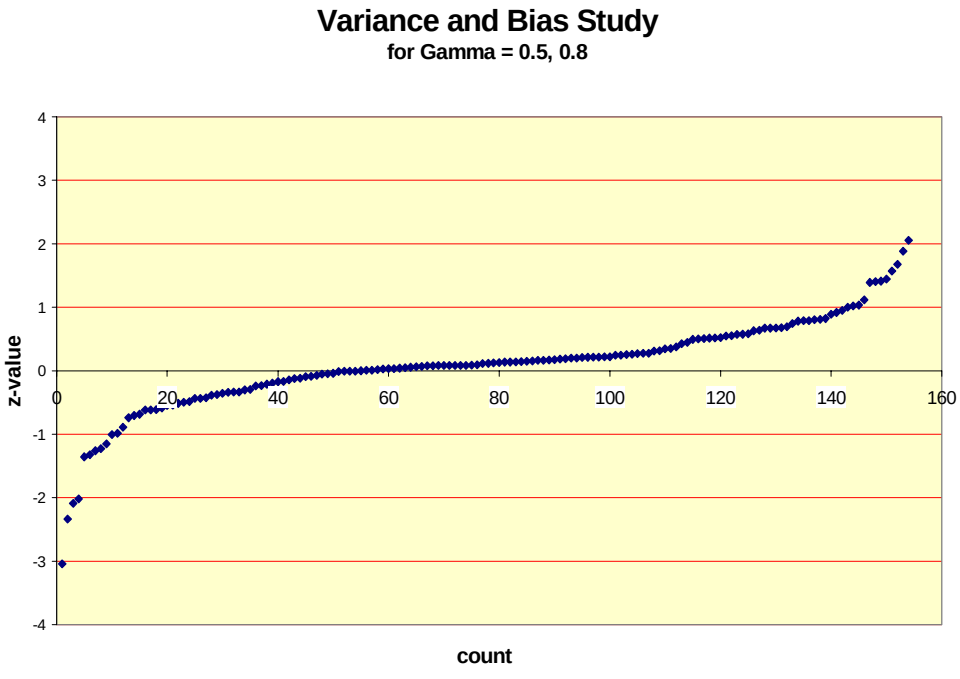
Histogram
for gamma = 0.5, 0.8 cases





smallest z'-values are -6.13 and -4.38

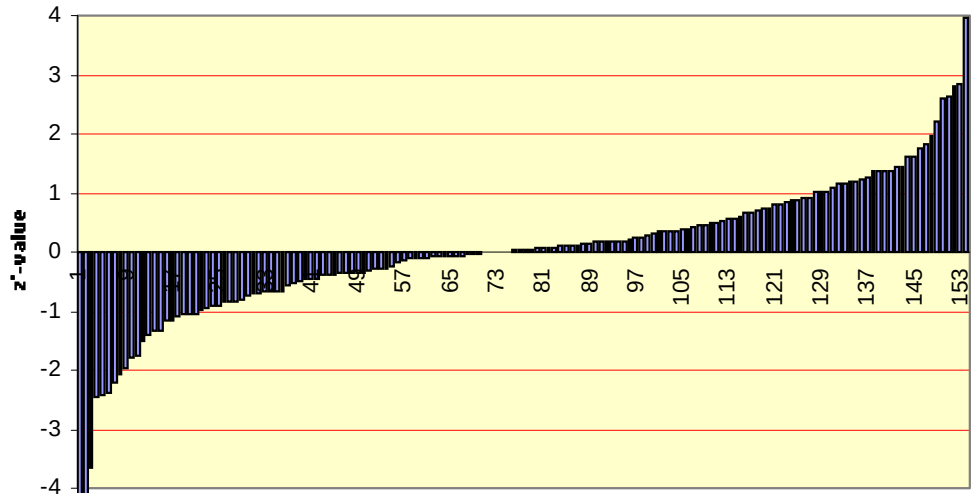
Possibly nonsampling error. Greater use of graphical edits (scatterplots) should prevent this.



Variance and Bias Study

for Various Gamma with Additional Bias Correction

Note: Each bar below points to one z'-value.

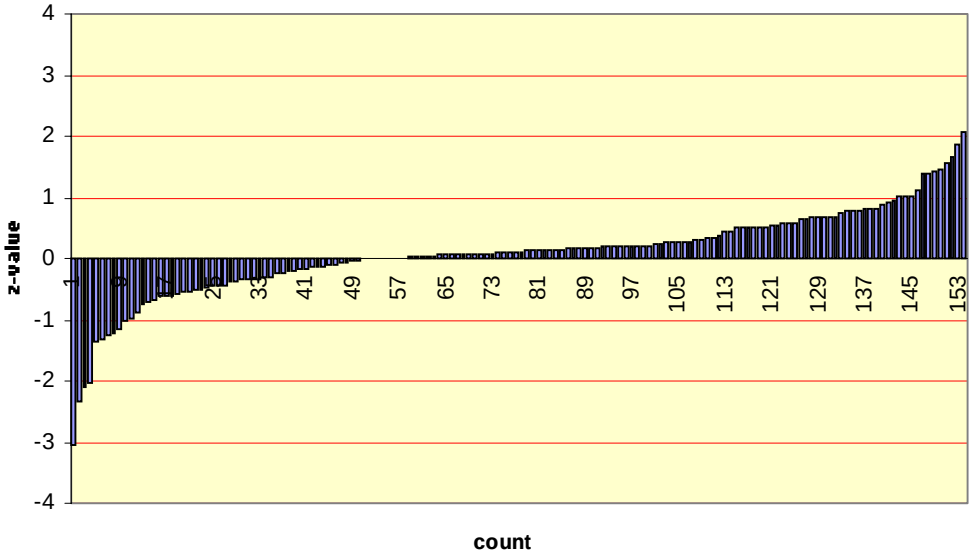


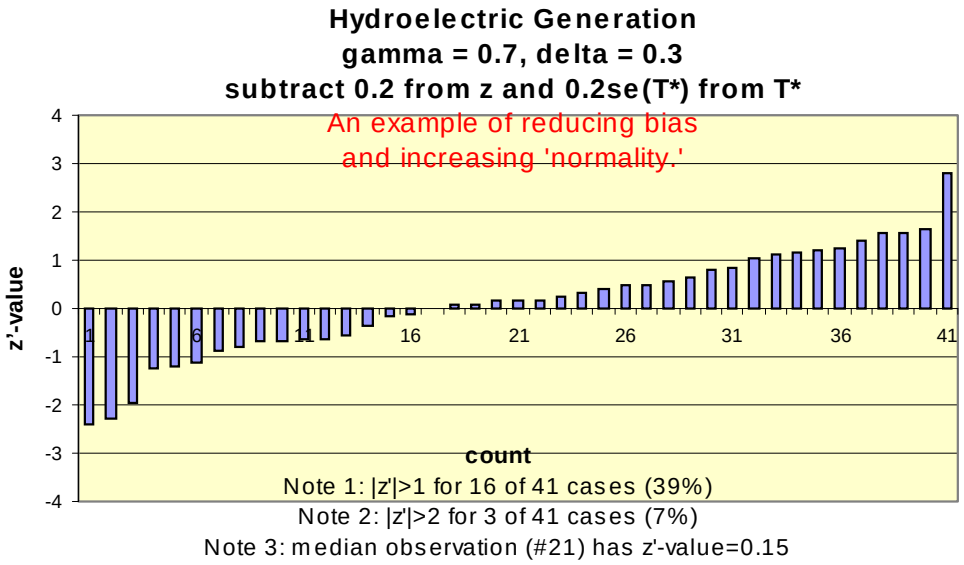
smallest z'-values are -6.13 and -4.38

Possibly nonsampling error. Greater use of graphical edits should prevent this.

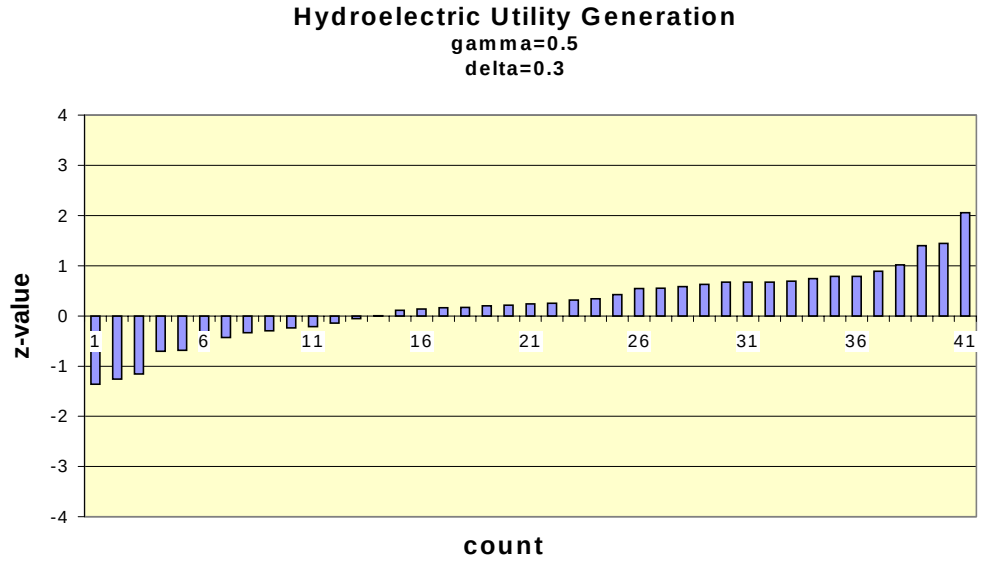
Variance and Bias Study

for Gamma = 0.5, 0.8

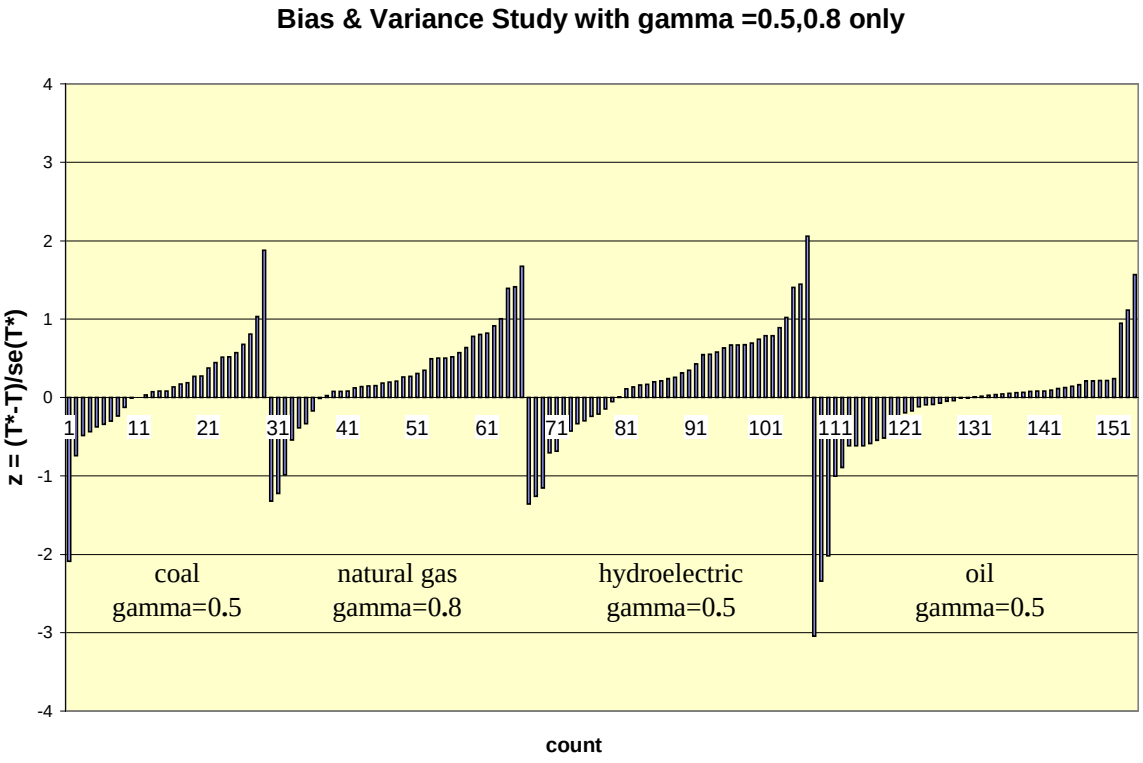




Note 1: |z'|>1 for 16 of 41 cases (39%)
Note 2: |z'|>2 for 3 of 41 cases (7%)
Note 3: median observation (#21) has z'-value=0.15



Note 1: |z| > 1 for only 7 of 41 cases (17%)
Note 2: median observation (#21) has z-value = 0.24



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 some overestimation 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, using monthly sampled observations, may proceed more smoothly. The ratio estimate (gamma=0.5) in particular, has shown robust behavior when estimating totals and will not often underestimate variance. 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: Stratified Variances Explained:

An example of a "partial" data file illustrating the new method is found on page 8 of Knaub(1999). For purposes of demonstration, consider a slightly modified version of this file to be a data set for a very small universe, which is shown below. Lower case letters are used to identify data for individual members of the universe. The "EG" designations identify strata within a "PG," where "EG" means "estimation group," and "PG" means "publication group."

y_i or y_i^*	$S1_i$	$S2_i$	EG	PG1	PG2
a) 6725	0	0	1	1	2
b) 6114	0	0	1	2	3
c) 5822	0	0	2	1	2
d) 4359	0	0	1	2	1
e) 3944	0	0	2	1	2
f) 2231	0	0	1	1	3
g) 1289	0	0	2	1	1
h) 1005	0	0	2	2	1
i) 892	0	0	2	2	3
j) 497	20	17	1	1	3
k) 455	18	16	2	2	1
l) 317	13	11	1	2	2
m) 295	12	10	1	2	1
n) 278	10	9	1	1	3
o) 246	10	9	2	1	2
p) 223	9	8	2	1	3
q) 211	8	7	1	1	1
r) 189	6	5	1	1	3
s) 181	6	5	2	2	1
t) 173	6	5	2	1	3
u) 141	5	4	1	1	3

Therefore, the data are from a universe of size N = 21, with n = 9 respondents (a through i) in a stratified sample. Ignoring nonsampling error (which is discussed in the above referenced article), values of zero for S1 and S2 indicate that the value in the first column is an observation, or y_i value, but positive values for S1 and S2 indicate that data are for imputed values in the first column, y_i^*.

Here, $S1_i^2 = \mathrm{V}_{\mathrm{L}}^{*}(y_i^{*} - y_i)$, which is the square of STDI in SAS, and $S2_i^2 = \sigma_e^{*2} / w_i$.

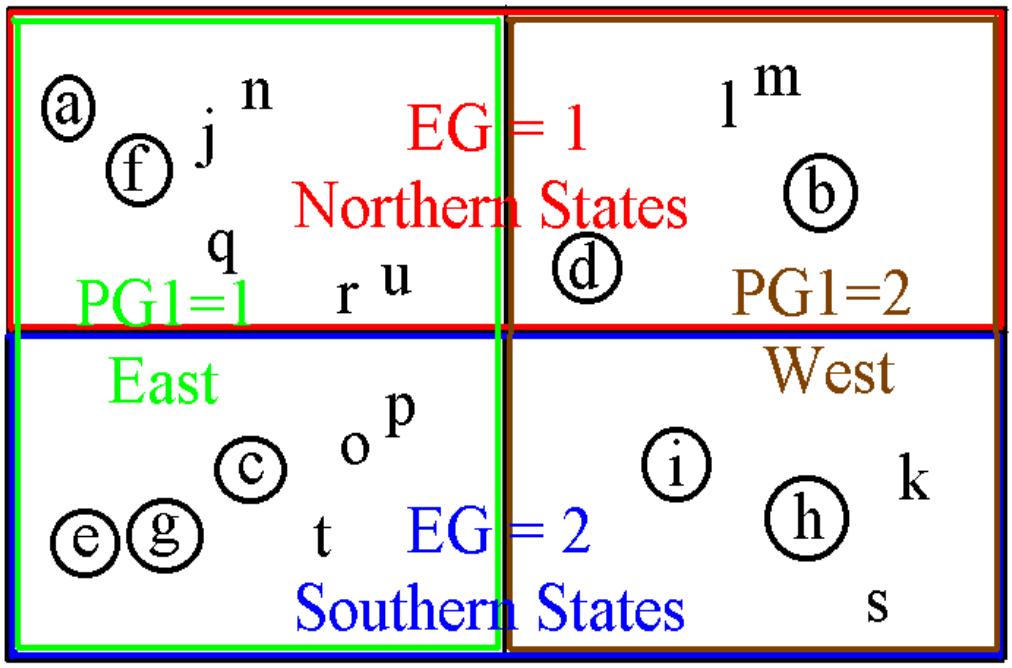
In general, from Knaub(2000), and as shown earlier:

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

$0 < \delta < 1$. $\mathrm{V}_{\mathrm{L}}^{*}(\mathrm{T}^{*} - \mathrm{T})$ is robust when using $\delta = 0.3.$

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

To demonstrate this with the above data table, consider, for example, that the EG are the geographic regions, North and South, which could be an important weather related factor for purposes of modeling. Thus EG = 1 could represent northern States, and EG = 2 could represent southern States. Suppose that PG1 is a publication grouping scheme, also based on geography, but perhaps there may be a need to publish totals (and standard errors) for eastern as opposed to western States. Thus, PG1 = 1 could represent eastern States, and PG1 = 2 could represent western States. (Other publication grouping schemes, PG2, etc., would be possible, but only one estimation grouping, EG, would have been used.) Thus, the following figure could demonstrate groupings for the lettered data above:



The data points that are available for modeling are circled. Those which had to be imputed are <u>not</u> circled. Estimated totals are easy. 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. However, to estimate standard error so that we may judge the accuracy of the total, attention must be paid to the estimation group strata used. Note then that the estimated variance for the total in the western States (PG1 = 2) is the sum of the estimated variance for the part of PG1 = 2 that is in the north (where EG = 1), and the estimated variance for the part of PG1 = 2 that is in the south (where EG = 2). Thus, the variance estimate for the northern strata of PG1 = 2 sums over two points ($\sum_r$), l and m, using data in a, f, d and b in a model applicable to EG = 1. (Imputed values, or "predictions," are obtained for l and m from the same model and data.) The estimated variance for the southern strata of PG1 = 2 sums over k and s, using data from e, g, c, *i* and h in a model applicable to EG = 2.

A review follows using the western States as an example. To estimate the total for the data element of interest, add the observed values from d, b, *i* and h, and the imputed values from l, m, s and k. To estimate the standard error of that estimated total, first estimate the variance of the northern part of the western States by obtaining S1 and S2 information on l and m, using a model on a, f, d and b, making use of $V^*_L(T^* - T)$. Then estimate the variance of the southern part of the western States by obtaining S1 and S2 information on s and k, using a model on e, g, c, *i* and h, and calculating $V^*_L(T^* - T)$. Add those two variances, say $V^*_{L,\ PG1=2,\ EG=1}(T^* - T)$ and $V^*_{L,\ PG1=2,\ EG=2}(T^* - T)$, and then take the square root of that sum to obtain the estimated standard error for the estimated total. (Note that SAS PROC REG, or similar software, is used to impute numbers that contribute to the total, and the same model exercises also produce the S1 and S2 values used in the variance estimates.)

For very small values of N-n, such as in this example, δ could be substantially larger than the usual 0.2 or 0.3. However, for these smaller values of N-n, the impact due to δ on the estimated total variance becomes smaller. Using $\delta = 0.5$ here, for the first stratum in the example above, applying the data in the above table corresponding to the above figure:

$$V^*_L(T^* - T) = 0.5(2)((13*13 - 11*11) + (12*12 - 10*10)) \quad + \quad (11*11 + 10*10)$$
$$= 92 + 221 = 313$$

(As a note on sensitivity, consider that for $\delta = 0.8$, the estimated variance for this stratum would be 368. The small impact on standard error is shown below.)

For the other stratum involving s and k, the estimated variance for $\delta = 0.5$ is 360, but would be 407 if $\delta = 0.8$. For $\delta = 0.5$, the standard error would be the square root of 313 + 360, or approximately 26. For the case of $\delta = 0.8$, the estimated standard error of the total would be approximately 28. The estimated total is 13618, of which 12370 was observed, and the remaining 1248 is the total of the imputed values. The estimated relative standard error is (26/13618)100%, or about 0.19%. The estimate to be published should then be 13600, or at best, 13620.

Epilogue:
This method is now being implemented for two sample surveys, and may be tested as an imputation method for one census survey in the near future, possibly to be expanded to several others. Current test data results are 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).) It also has promise as a small area technique.

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 in the course of model applications. A 'point-and-click' version of these scatterplot edits should be a great help to the data managers.

Acknowledgments:
Thanks to the American Statistical Association's Committee on Energy Statistics for comments that at 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 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. (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, 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," to appear 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. (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 2000.)

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.

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.

Appendix:
Extract from a SAS program used to implement this method:

Code below was extracted from a SAS program written by Dr. Orhan M. Yildiz, an analyst contracted to apply this method to Energy Information Administration (EIA) surveys. Fragments of this code may be useful to various readers, so it is included here.

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 that establishment. 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 are labeled "ADD-ONs" here. (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" would not be included in such graphs. PROC PLOT below is used for such data editing, although a "point-and-click" version of these scatterplot graphics would be preferable.) 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. Other parts of the code below make it more convenient to specify the number of regressors to be used for various sectors of the universe, specify different regressor weights, etc. Although this code is part of a program dealing with a specific survey, it is shown here as an aid for other statisticians/programmers/analysts in implementing this methodology for other surveys. (Also see the shorter, more generalized code on pages 33 and 34 of Knaub(1999).)

```
DATA ANDAT (DROP=AC);
  SET &RTY.TDAT;
* SELECTION OF FINAL DATA FOR THE MODEL;

* EXCLUDE ADD-ONs FROM THE MODEL.;
  IF AOFLAG='AO' THEN DELETE;

* ANALYST CONTROLS THE VALUE OF GAMMA.;
  IF FSRCE='NGAS' THEN GAMMA=0.8;
  ELSE                 GAMMA=&GAMMA;

  PROC SORT DATA=ANDAT;
    BY FSRCE STRATA;

  PROC PLOT DATA=ANDAT;
    BY FSRCE STRATA;
      PLOT Y*X1M      / HZERO VZERO;
      TITLE "&RTY &MOD vs. &MOD.(T-1)/12 &
           &VX2 for DATE: &MDATE.";
```

```
%MACRO mcRPARTS(SPLT);
 %IF &SPLT=ONE %THEN %DO;
  DATA ANDAT;
   SET ANDAT;

   IF CHECK='YES' THEN W=((X1M+X2)**(-2.0))**GAMMA;
   ELSE                W=((X1M   )**(-2.0))**GAMMA;

   PROC REG DATA=ANDAT OUTEST=EPARAM EDF;
    MOD1: MODEL Y= X1M &VX2/NOINT MSE;
    BY FSRCE STRATA;
    WEIGHT W;
    OUTPUT OUT=REGDAT P=YP R=RS
    STDP=SMPY STDR=SRES STDI=SIPY;
 %END;
 %ELSE %IF &SPLT=TWO %THEN %DO;
  DATA ANDAT1 ANDAT2;
   SET ANDAT;

   IF FSRCE='Coke' OR
    (FSRCE='Hydroelectric' & STRATA='S') THEN OUTPUT ANDAT1;
   ELSE                                OUTPUT ANDAT2;

   DATA ANDAT1;
    SET ANDAT1;

   W=((X1M   )**(-2.0))**GAMMA;

   PROC REG DATA=ANDAT1 OUTEST=EPARAM EDF;
    MOD1: MODEL Y= X1M/NOINT MSE;
    BY FSRCE STRATA;
    WEIGHT W;
    OUTPUT OUT=REGDAT1 P=YP R=RS
    STDP=SMPY STDR=SRES STDI=SIPY;

  DATA ANDAT2;
   SET ANDAT2;

   W=((X1M+X2)**(-2.0))**GAMMA;

   PROC REG DATA=ANDAT2 OUTEST=EPARAM EDF;
    MOD1: MODEL Y= X1M &VX2/NOINT MSE;
    BY FSRCE STRATA;
    WEIGHT W;
    OUTPUT OUT=REGDAT2 P=YP R=RS
    STDP=SMPY STDR=SRES STDI=SIPY;

  DATA REGDAT;
   SET REGDAT1 REGDAT2;

  PROC DELETE DATA=REGDAT1 REGDAT2;
 %END;
```

```
%MEND mcRPARTS;

%mcRPARTS(TWO);

PROC SORT DATA=REGDAT;

  BY FSRCE STRATA;

PROC DELETE DATA=ANDAT;
*_____________________________________________*
*                                             *
*        RELATIVE STANDARD ERRORS             *
*_____________________________________________*;

DATA EPARAM (KEEP=FSRCE STRATA MSQE EDF MDF);
  SET EPARAM;
* ASSIGN MSE FROM ESTIMATED PARAMETERS DATA;

  MSQE=_MSE_;
  EDF=_EDF_;
  MDF=_P_;


  PROC SORT DATA=EPARAM;
    BY FSRCE STRATA;

  PROC PRINT DATA=EPARAM;
    TITLE 'MEAN SQUARE ERROR BY STRATA';

DATA EREGDAT (KEEP=MRGCODE MSQE EDF MDF W YP SIPY);
* DATA EREGDAT (DROP=GAMMA SMPY SRES);
  MERGE REGDAT (IN=A) EPARAM (IN=B);
  BY FSRCE STRATA;
* MERGE REGRESSION DATA WITH PARAMETER ESTIMATES;

  IF (A & B);

PROC DELETE DATA=REGDAT EPARAM;

  PROC SORT DATA=EREGDAT;
    BY MRGCODE;

DATA AGGDAT;
  MERGE EREGDAT (IN=A) &RTY.TDAT (IN=B);
* BY BCODE FSRCE;
  BY MRGCODE;
* MERGE REGRESSION OUTPUT AND INPUT DATA WITH ADD-ONS.;

  IF B;

* ADJUST ADD-ON FLAG FOR THE ANALYSIS VARIABLE:
*   PR-PREDICTED, AO-ACTUAL, MX-MISSSING PREDICTOR, MA-MISG. ACT.
```

```
*   MM-MISSING.
*****************************************************************;

* IF  AOFLAG NE 'AO' & (FRFLAG='A' & Y=.);
  IF  AOFLAG NE 'AO' & Y=.
      THEN DO; YO=YP;
            S1=SIPY;
            V1=SIPY**2;
            V2=MSQE/W;

            S2=SQRT(V2);
            AOFLAG='PR';
            IF YO=. THEN DO; AOFLAG='MP';
                        YO=Y;
                     END;
          END;
  ELSE      DO; YO=Y;
            S1=.;
            V1=.;
            V2=.;
            S2=.;
            IF AOFLAG NE 'AO'
              THEN DO; AOFLAG='AC';
                    IF YO=. THEN AOFLAG='MA';
                  END;
          END;

PROC DELETE DATA=EREGDAT &RTY.TDAT;

  PROC SORT DATA=AGGDAT;
    BY FSRCP CENR STATE FSRCE;

* PRODUCE SUMMARY TABLES FOR TOTALS;
  PROC MEANS DATA=AGGDAT NOPRINT N NWAY MISSING NMISS SUM MIN MAX;
    CLASS FSRCP CENR STATE;
    VAR YO X1M;
    OUTPUT OUT=STATESUM SUM(YO X1M)=STATEYO STATEX1M;
    TITLE 'AGRREGATES BY PUBLICATION FUEL TYPE AND STATE';

  PROC SORT DATA=STATESUM;
    BY FSRCP STATE;

* SECTION TO OBTAIN PUBLICATION GROUP TOTALS.;

%MACRO mcSUMS(RGN);
  PROC SORT DATA=AGGDAT;
    BY FSRCP &RGN;

* PRODUCE SUMMARY TABLES FOR TOTALS;
  PROC MEANS DATA=AGGDAT NOPRINT N NWAY MISSING NMISS SUM MIN MAX;
    CLASS FSRCP &RGN;
    VAR YO X1M;
```

```
   OUTPUT OUT=&RGN.SUM SUM(YO X1M)=&RGN.YO &RGN.X1M;
   TITLE "AGRREGATE BY PUBLICATION FUEL TYPE AND &RGN REGION";

  PROC SORT DATA=&RGN.SUM;
    BY FSRCP &RGN;
%MEND mcSUMS;

%mcSUMS(CENR);
%mcSUMS(USA);

* SECTION TO OBTAIN PUBLICATION GROUP VARIANCES FOR IMPUTED DATA.;
DATA IMPDAT;
  SET AGGDAT;

* SELECT ONLY THE ANNUAL OBSERVATIONS FOR WHICH THERE IS IMPTN.;
  IF AOFLAG NE 'AO';
* IF FRFLAG='A' & Y=.;
  IF Y=.;

%MACRO mcPBGVAR(RGN);
PROC SORT DATA=IMPDAT;
   BY FSRCP &RGN FSRCE STRATA;

DATA ESTGVAR;
  SET IMPDAT;
  BY FSRCP &RGN FSRCE STRATA;

  DELTA=&DELTA;

  IF FIRST.STRATA OR FIRST.FSRCE
    THEN DO; XS=0;
           V1S=0;
           V2S=0;
         END;

  XS+1;
  V1S+V1;
  V2S+V2;

  IF LAST.STRATA OR LAST.FSRCE
    THEN DO; XN=XS;
           V1T=V1S;
           V2T=V2S;
           VT=DELTA*XN*V1T+(1.0-(DELTA*XN))*V2T;
           OUTPUT ESTGVAR;
         END;

PROC SORT DATA=ESTGVAR;
  BY FSRCP &RGN;

DATA &RGN.VAR;
  SET ESTGVAR;
```

```
  BY FSRCP &RGN;

  IF FIRST.&RGN THEN VATS=0;
  VATS+VT;

  IF LAST.&RGN THEN DO; VARYO=VATS;
                  OUTPUT &RGN.VAR;
                END;
%MEND mcPBGVAR;

%mcPBGVAR(STATE);
%mcPBGVAR(CENR);
%mcPBGVAR(USA);
```