Using Prediction-Oriented Software for Model-Based and Small Area Estimation

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

KEY WORDS: survey sampling; prediction; estimation; imputation; model variance estimation

ABSTRACT: Survey sampling and inference may be accomplished by solely design-based procedures, or solely model-based procedures, or by model-assisted, design-based procedures. Depending upon circumstances, there are advantages to each of these methods. There are times, particularly in (highly skewed) establishment surveys, when, either in terms of resources, and/or nonsampling error, it may not be practical to sample from among the 'smallest' members of the population, and solely model-based procedures may then be advantageous. This paper shows a general approach that may be used to organize such estimations in a flexible manner. Readily available regression software may be used and results may be easily reorganized to present various aggregations. (Note: See Knaub (1999a) for an expanded version of this paper with greater use of examples and a discussion of the usefulness of this method as an imputation procedure.)

BACKGROUND:

Cutoff model-based sampling has proven to be useful at the Energy Information Administration (EIA) for establishment surveys because of practical problems that arise when trying to sample from among the smaller members of a population. These problems involve both timeliness and nonsampling error. An obvious example of this occurred when it was observed that a small electric utility only read its meters once every three months. A census of electricity sales is performed annually, but a sample is done monthly. To ask that utility to participate in a monthly sample would not be very practical.

Since early discussions by Brewer (1963) and Royall (1970), and even comments by Cochran (1953), mentioned in Knaub (1995), most model-based sampling has probably only made use of simple linear regression, with a fixed zero intercept. (We will use a common misnomer and say "no intercept.") At the opposite extreme, econometric applications of regression for prediction have often used more intricate models, although perhaps too often they are overspecified. (Note here that Carroll and Rupert (1988) is an excellent monograph, with great application to model-based inference as well as many other applications.)

Royall and Cumberland (1981) studied improvement in variance estimation for the simple linear regression model, but, from Knaub (1992), page 879, Figure 1, it can be seen that in practice, improvement, if any, may be negligible. Royall has also considered the incorporation of randomization in the sampling procedure, and seems to have moved away from cutoff sampling. Brewer (1995) showed how model-based sampling and inference, and design-based sampling and inference may complement each other. Further, model-assisted, design-based sampling (see Sarndal, Swensson and Wretman (1992), Chaudhuri and Stenger (1992), *et. al.*) has become fairly popular. Other works, such as Sweet and Sigman (1995), and Steel and Fay (1995), have also advanced the use of models in a supporting role. However, for purposes of imputation and/or cutoff sampling, model-based applications may have been given too little emphasis. Cutoff sampling can be very useful for highly skewed establishment surveys, not only because of practical data collection problems, but because of the efficient use it makes of resources. (See Royall (1970).) This should not be ignored.

Although the simple linear regression model is very useful for inference from model-based survey sampling, sometimes a multiple linear regression model may be more useful. In Knaub (1996), we see an example. (This is expanded upon in Knaub (1997).) One may use test data to see which model (with one, two or more regressors) performs best, but there can be other considerations also. In Knaub (1996), the variate of interest was electricity sales for resale among a certain type of generators. The first regressor used was the same variate from a previous census. However, it was found that a lot of those generators were sporadic in their sales for resale, sometimes using all electricity for their own purposes, or perhaps not producing electricity for extended periods. It was not uncommon to have non-zero current sales for resale, but a zero value for the corresponding regressor. However, when a second regressor, generating capacity, was introduced, the situation was much improved. Every generating plant must have a positive capacity value. If this procedure were not used, then all cases where a zero value for sales for resale was recorded in the previous census would need to be handled separately, perhaps as a separate stratum within the sample. A census or a design-based sample could be performed within that stratum.

NEW METHODOLOGY:

For each of several models (with or without an intercept, and with one or two regressors), the author wrote test computer programs at the Energy Information Administration (EIA) to estimate totals and relative standard errors of the estimated totals, for multiple applications. It became obvious that it would be advantageous to utilize existing vendor generated programming if the future held possibilities of using more regressors, or at least the need to test and consider alternatives. The SAS system is available to EIA employees, so the use of SAS PROC REG was explored. This, however, is not to be considered an official endorsement of SAS products. In fact, **any statistical software package that will provide (1) predicted values, (2) a standard error or variance of the prediction error, and (3) the mean square error (MSE) from the analysis of variance, will suffice. Thus, this paper relates to any prediction-oriented software.** Such software will provide predictions, using a specified model, for every specified member of a set of potential respondents from whom data were not collected, provided that relevant data are collected. If we add those predictions and the collected values, then we obtain exactly the same estimated total that we would have obtained using the more traditional model-based approach to inference, if the set of data used is from the population or the portion of a population for which we seek to estimate a total. The estimation of variance will differ, but this was studied and an example will follow. For now, however, note the implications of this new organization of the estimation procedure. We need not limit ourselves to sampling within a category whose total we wish to estimate! When the time comes to present an estimated total for a given category, we need only to have a collected or a predicted value for every member of the population falling within that category. We then simply add the appropriate collected and predicted numbers to obtain the estimated total. Each member will have a value for the variate of interest. If the value was not collected, then ideally it will have been estimated from the optimal regression model using the optimal corresponding sample. The data in that sample need not have been limited to the category for which we want an estimated total. The "optimal sample," in each case, would be one in which there is a compromise between sample size and heterogeneity. That is, using a larger sample by including a broader group of respondents only helps if these respondents do behave similarly under the model. (That is, the model and the parameter values chosen must relate reasonably well to all data to which the model is applied.)

ESTIMATION OF VARIANCE:

Here we will use $V_L^*(T^* - T)$ to represent the variance of an estimated total (or the variance of the error in estimating the total). This is a multivariate extension of V_L found in Royall and Cumberland (1981).

Here $V_L^*(T^* - T) = \sum_r \sigma_e^{*2} / w_i +$

$(N-n)^2 \, V^*(b_0) + (\sum_r x_{1i})^2 V^*(b_1)$

$+ (\sum_r x_{2i})^2 V^*(b_2) + (\sum_r x_{3i})^2 V^*(b_3)$

$+ \ldots + 2(N-n)(\sum_r x_{1i}) \mathrm{COV}^*(b_0, b_1)$

$+ 2(\sum_r x_{1i})(\sum_r x_{2i}) \mathrm{COV}^*(b_1, b_2) + \ldots$

where $\sum_r$ means to sum over the cases not in the sample (Royall (1970)). 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; $(N - n)$ is the number of members of the population that are not in the sample; the b's are regression coefficients; and the x's are regressors. V^* and COV^* are estimates of variance and covariance.

For the case where the data element of interest is collected for all members of a population, but we wish to 'predict' a value for a new case, then the variance of the prediction error is represented by $V_L^*(y_i^* - y_i)$. (See Maddala (1992).) This may usually be considered to be a way to predict "future observations" (Maddala (1977), page 464), but it could be used to estimate for a single 'missing' observation. Note that when $N - n = 1$, then $V_L^*(T^* - T) = V_L^*(y_i^* - y_i)$. Now consider the form taken by the sum over r of $V_L^*(y_i^* - y_i)$:

$\sum_r V_L^*(y_i^* - y_i) = \sum_r \sigma_e^{*2} / w_i + (N-n) V^*(b_0)$

$+ (\sum_r x_{1i}^2) V^*(b_1) + (\sum_r x_{2i}^2) V^*(b_2)$

$+ (\sum_r x_{3i}^2) V^*(b_3) + \ldots + 2(\sum_r x_{1i}) \mathrm{COV}^*(b_o, b_1)$

$+ 2(\sum_r x_{1i} x_{2i}) \mathrm{COV}^*(b_1, b_2) + \ldots$

Now consider regressors, x_j, where for each regressor, the values are fairly constant. In the extreme, if x_{ji} is a constant for all i (but may be different for j), then we have, summing over i for a given j,

$(\sum_r x_{ji})^2 = (c_j \sum_r 1)^2 = c_j^2 (N-n)^2$ and

$\sum_r x_{ji}^2 = c_j^2 (N-n)$, so in the extreme,

$(\sum_r x_{ji})^2 / \sum_r x_{ji}^2$ approaches $N-n$.

Similarly, in the extreme case:

$(\sum_r x_{ki})(\sum_r x_{li}) = c_k c_l (N-n)^2$ and

$\sum_r x_{ki} x_{li} = c_k c_l (N-n)$.

Therefore, with $0 < \delta < 1$, approximate as follows:

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

Although $V_L^*(y_i^* - y_i)$ and σ_e^{*2}/w_i are usually nearly equal in many practical applications **(the difference being negligible and not considered in Knaub(1998), where the coefficients were dealt with as if they were constants)**, there is a cumulative impact here that is not negligible when $(N-n)$ becomes somewhat larger than 1. Further, the nature of δ was explored. (See Knaub (1999a) for a discussion of this topic.) Using both real and artificial test data, it appears that for establishment surveys, δ = 0.3 might be used within strata, and then the variances corresponding to each of the strata may be added. Stratification should reduce the variances of the parameters and make δ less important to the estimate of variance of the estimated total. (For household surveys, the optimal δ may be smaller, perhaps 0.2, as mentioned in Knaub (1999a).)

ADVANTAGES OF NEW METHODOLOGY:

One advantage is that data can be estimated using the most efficient groupings available. If that data grouping is just the category for which we are trying to estimate a total, then we will obtain the same estimated totals as when using Brewer (1963), Royall (1970), Knaub (1996), and others. (Standard errors will be slightly different due to the approximation above.) However, this method easily allows the use of any data set one chooses to designate for purposes of predicting each 'missing' number. (A number is 'missing' if it was not collected/observed. This could be a number for an entity not in a sample, or for a nonrespondent to a sample or a census.) The larger and more homogeneous the set of data used in each case, the better the predictions, and the better the overall estimation of the total. Thus, this is a more powerful method than if we were to only consider 'borrowing strength,' a common term in small area estimation, where one may use data from a 'neighboring' area when the data in the area for which one wishes to report are too sparse.

Another advantage, the one for which this method was created, is that the model can be quickly altered when necessary. That is, regressors may be added or deleted, as well as the intercept term, and the regression weight may be easily altered.

In practice, this means a file is to be built containing collected values, predicted values, and values for $V_L^*(y_i^* - y_i)$, and σ_e^{*2}/w_i. Each record of the file will therefore contain either an observed or a predicted number, two variance related numbers (each set to zero if there is an observed number), and indicators to identify the groupings used to estimate the predicted numbers, and to identify possible categories for which we may wish to estimate subtotals. This yields a highly organized and flexible file that can satisfy the customer who wants to see what was added to obtain a given subtotal, and can be archived and later examined without ambiguity. (EIA customers have asked for this kind of information.) Further, later regrouping of data will be easy. Recently, for example, the boundary between two of the North American Electric Reliability Council (NERC) Regions changed substantially due to several companies changing their affiliations. Accounting for such a change would be easy when using this new methodology.

Consider a typical data file 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. A few rows (records) of such a file could appear as follows:

y_i or y_i^*	$S1_i$	$S2_i$	EG	PG1	PG2
4359	0	0	1	2	1
497	20	17	1	1	3
317	13	11	1	2	2

Next, suppose that we have some information on nonsampling error. Although nonsampling error is difficult to measure, tables of revision 'errors' (*i.e.*, changes made) are sometimes maintained. The relative percent change between preliminary and final submissions from respondents may give some indication of the severity of nonsampling error. The S1 and S2 values would be impacted by nonsampling error. In spite of the lack of information and the complicated nature of the true relationships between errors, it may be instructive to perform a data quality study occasionally, that would supplement the S1 and S2 values above so that applying the variance formula would no longer estimate only model variance, but instead would, to an extent, approximate overall error. For example, in the partial table above, we might, based on revisions, estimate that $d_i \approx 0.02y_i + 3$, if the d_i are to replace the zero values associated with each observed value, y_i. The S1 and S2 values above might also be replaced by $r_i \approx \left[\left(0.02y_i^*\right)^2 + S_i^2\right]^{1/2}$, using "S" in place of either "S1" or "S2." This would yield the following:

y_i or y_i^*	d_i or r_{1i}	d_i or r_{2i}	EG	PG1	PG2
4359	90	90	1	2	1
497	22	20	1	1	3
317	14	13	1	2	2

EXAMPLE (using hydroelectric generation):
This example is for hydroelectric generation in the Western United States. Logically, it would seem that US hydroelectric generation may best be collected within the US Standard Regions for Temperature and Precipitation used by the National Climatic Data Center of the National Oceanic and Atmospheric Administration (NCDC/NOAA) since one would like to have data grouped as homogeneously as possible for the largest data set possible. When estimating (technically, 'predicting') hydroelectric generation for 'missing' observations (*i.e.*, for that part of the population that was not on the sample, or, whether sample or census, did not respond), the regressors used here were generator capacity and previously reported generation from a census. The relationship of those numbers to a current sample or census of generation might be expected to differ by geographic region because the changes in precipitation might be expected to be similar within the NCDC regions. However, the Energy Information Administration (EIA) would publish such generation numbers by Census division (originally from the Bureau of the Census) or NERC region (from the North American Electric Reliability Council). Maps are included in Knaub (1999a) for all of these regions. The EIA might also publish these numbers by State.

There is an NCDC region in the West consisting of California and Nevada, and another region north of that consisting of Oregon, Washington and Idaho. However, California, Oregon and Washington are in what EIA refers to as the Pacific Contiguous Census Division.

What if hydroelectric generation data were collected in each of the two NCDC regions just mentioned, and data 'predicted' for all members of the population for which data were not collected in those regions, and then a total was published across the Census subdivision for California, Oregon and Washington? An example will be shown to illustrate cutoff model-based sampling and inference. (Imputation is very similar, as shown in Knaub (1999a), and presented in Knaub (1999b).)

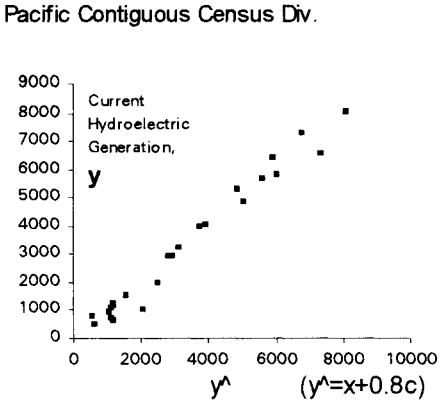
In the following example, the "cutoff" will be that data are not collected from hydroelectric plants with less than 200 megawatts of generating capacity. These data were chosen for use in testing as they represent a reasonable set of real data to isolate for this purpose. The author employed a testing technique used for a number of years since being suggested by Dean Fennell at the EIA. The data in this case come from two annual censuses. Data are removed from one census to simulate a sample with the remaining data. The other census supplies regressor data. After 'prediction'/estimation has been accomplished, the results are compared to what had been collected before the artificial 'sample' was formed. With enough such test data sets, one may judge the

performance of both total estimation and variance estimation to some degree.

For the NCDC/NOAA West region, there are only 7 hydroelectric plants out of 231 that meet the 200 megawatt capacity threshold. They account for just over 20% of the generation. Two regressors are used: generation from a past census, x, and plant capacity, c. A preliminary estimate of y, $\hat{y} = 0.5x + 1.3c$, was used in the regression weight. Regression weights for these models are discussed in Knaub (1999a).

For the NCDC/NOAA Northwest region, there are 25 hydroelectric plants out of 147 that meet the 200 megawatt capacity threshold. They account for about 85% of the generation. The same two regressors are used: generation from a past census, x, and plant capacity, c. A preliminary estimate of y, $\hat{y} = x + 0.9c$, was used in the regression weight.

For the Pacific Contiguous Census Division, there are 28 hydroelectric plants out of 331 that meet the 200 megawatt capacity threshold. They account for about 70% of the generation. Two regressors are used: generation from a past census, x, and plant nameplate capacity, c. The sample (n = 28) of generation values, y, is plotted against $\hat{y} = x + 0.8c$ below. Once again, $\hat{y}$ was a preliminary estimate of y, to be used in the regression weight.



When **the Pacific Contiguous Census Division region** (CA, OR and WA) is used as a category from which to sample, the model does not perform as well. Although graphs for these data can be somewhat deceiving, the graph above indicates that heteroscedasticity is not a well-defined phenomenon in the case of the Pacific Contiguous Census Division for these data. The sample size is small, but graphs showing most of the hydroelectric generation data in this region are found in Knaub (1999a), and they also appear a little odd.

$y_i = \beta_x x_i + \beta_c c_i + e_0 (x + 0.8c)^{\gamma}$, with $\gamma = 0.8$, is the model used here, as discussed in Knaub (1999a), which contains a section on regression weights with references to more information. Further, we have:
R-square: 0.978
x coefficient: 1.01; standard error: 0.08
c coefficient: 0.75; standard error: 0.29
n = 28; N = 331; T = 188,710 gigawatthours (GWh)
T^* = 200,681 and $T - T^*$ = -11,971.

Performance appears to be degraded as a result of sampling from within a category (the Pacific Contiguous Census Division) that is too heterogeneous.

However, after making the predictions by NCDC region in accordance with the new method, the predicted and collected data were aggregated within the Pacific Contiguous region, by NCDC regions. For each NCDC stratum, we use $\delta = 0.3$ and $\gamma = 0.8$, as mentioned in Knaub (1999a). One then obtains $T^* = 184,597$, and therefore $T - T^*$ = 188,710 - 184,597 = 4113 with an estimated standard error of 5327. To estimate 189 terawatthours as 185 is decidedly better than 201, although this is only one example.

For household surveys, data are typically not as skewed as in establishment surveys, and data sets tend to be larger, whether one is dealing with sampling inference, imputation, or both. This paper, however, concentrates on the generally smaller, heavily skewed establishment survey, particularly where a cutoff sample is most practical. However, with possible adjustments to regression weights, and to δ, the general procedure should be very widely useful, especially for imputation.

SMALL AREA ESTIMATION:

When regressor data are available, and models can be used to predict a response for any member of the population, then an estimated subtotal may be produced for any subgroup. If we wished to estimate a hydroelectric generation total for a given State from which we had collected few if any responses (but we have complete regressor data), then we may do so. Accuracy may be too low to make the result useful, but this is always true for small area estimation. In this case, however, one can estimate a standard error, although

accuracy of that statistic, as well as the estimated total, would be dependent upon the accuracy of homogeneity assumptions. In the case of the above examples, it was found that data were homogeneous by State to the extent that for the States in the example, it would have been best to have sampled and estimated by State. Thus if all data were missing for a given State, it may be inappropriate to 'predict' each of those missing values using data from other States. A variance estimate would be produced, but it may not be acceptably accurate. However, if the estimated standard error is not too large for the purposes to which the data are to be used, **and one states the homogeneity assumptions being made**, it may be possible to provide such an estimated total to data 'customers' without misinforming them.

REFERENCES:

Brewer, K.R.W. (1963), "Ratio Estimation in Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process," Australian Journal of Statistics, 5, pp. 93-105.

Brewer, K.R.W. (1995), "Combining Design-Based and Model-Based Inference," Business Survey Methods, ed. by B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott, John Wiley & Sons, pp. 589-606.

Carroll, R.J., and Ruppert, D. (1988), Transformation and Weighting in Regression, Chapman & Hall.

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

Chaudhuri, A. and Stenger, H. (1992), Survey Sampling: Theory and Methods, Marcel Dekker, Inc.

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. (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/InterStat. (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/InterStat. (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 1997.)

Knaub, J.R., Jr. (1998), "Filling in the Gaps for A Partially Discontinued Data Series," InterStat, October 1998, http://interstat.stat.vt.edu/InterStat. (Note shorter, more recent version in ASA Business and Economic Statistics Section proceedings, 1998.)

Knaub, J.R., Jr. (1999a), "Using Prediction-Oriented Software for Survey Estimation," InterStat, August 1999, http://interstat.stat.vt.edu/InterStat.

Knaub, J.R., Jr. (1999b), "Using Prediction-Oriented Software for Estimation in the Presence of Nonresponse," presented at the 1999 International Conference on Survey Nonresponse, American Statistical Association.

Maddala, G.S. (1977), Econometrics, McGraw-Hill, Inc.

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

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.

Sarndal, C.-E., Swensson, B. and Wretman, J. (1992), Model Assisted Survey Sampling, Springer-Verlag.

Steel, P. and Fay, R.E. (1995), "Variance Estimation for Finite Populations with Imputed Data," Proceedings of the Section on Survey Research Methods, Vol. I, American Statistical Association, pp. 374-379.

Sweet, E.M. and Sigman, R.S. (1995), "Evaluation of Model-Assisted Procedures for Stratifying Skewed Populations Using Auxiliary Data," Proceedings of the Section on Survey Research Methods, Vol. I, American Statistical Association, pp. 491-496.