Quasi-Cutoff Sampling and Simple Small Area Estimation with Nonresponse
James R. Knaub, Jr.1

Abstract:
Here, small area estimation is applied in the sense that we are “borrowing strength” from data outside of given subpopulations for which we are to publish estimated totals, or means, or ratios of totals. We will consider estimated totals for establishment surveys. A subpopulation for which we wish to estimate a total will be called a “publication group” (PG), and data that may be modeled together, using one regression, will be called an “estimation group” (EG). See Knaub(1999, 2001, 2003) regarding this for a more complex application. When a PG consists of a set of EGs, that is stratification. When an EG contains PGs, this is a simple form of small area estimation because we are using data outside of a given publication group to help estimate statistics/parameters for that model, used to estimate for each impacted PG. (In Knaub(1999, 2001), there are overlapping ‘areas’ as well.) Here we consider very small areas (PGs), which may fall within a ‘larger’ EG, and here we are only considering one regressor, but this could be generalized. Sample sizes and population sizes considered in this paper can be very small within a given PG, say a State and economic end-use sector. In the case of \( n = N = 1 \), a single response is the total for that PG. If it is part of an EG with other data, then if there is a nonresponse in that case, an estimate in place of that observation may be obtained for contribution, for example, to a US-level aggregate number for that end-use sector, and a variance contribution to be added to the US-level variance would be found as well. Further, a scatterplot for such an estimation group, especially if a confidence band were constructed (Knaub(2009), section 4, and Knaub(2012b), Figure 1) could be used to help edit data. If that PG with \( n = N = 1 \) were looked at alone, one could not have a scatterplot that would determine if a response were reasonable for the current circumstances. (A forecast for that one point would not be as good if some event were to cause a break in the time series, and one would have to consider a time series for every single point, many more graphs, and for some there would be no series available. But a scatterplot to accompany this regression modeling would consider every point used in the model. Data for which there are no regressor data, such as “births,” are “added on” to totals outside of modeling.) Techniques here may be used for estimation (“prediction”) for sample surveys, and to impute for nonresponse for sample surveys and census surveys. There may be applications to other fields of statistics as well.

Key Words: Regression, Model-Based Estimation, Weighted Least Squares, Scatterplots, Small Area Estimation, Data Editing, Establishment Surveys, Seasonality, Borrowing Strength

1. Introduction

Here we consider the integration of sampling, estimation, and data analysis for establishment survey data with the possible use of small area estimation. The instrument which ties these three areas together — (1) model-based estimation, (2) sampling, and (3) data analyses — is scatterplot graphics, as will be illustrated below, in four figures. The regression here, or more complex techniques (Knaub(1999)), may have other applications, but will be presented here in terms of survey statistics, and establishment surveys in particular, where sample sizes \( n \) and population sizes \( N \) are small for this paper.

1 Lead Mathematical Statistician; Statistical Methods Team, US Energy Information Administration. - Disclaimer: The views expressed are those of the author, and are not official US Energy Information Administration (EIA) positions unless claimed to be in an official US Government document.
As perhaps sometimes the case with econometric data, we may just be using available data, or here, observations for the largest cases (measured by a single regressor or perhaps a linear combination of regressors as a ‘size’ measure (Knaub(2012a)). If only the largest cases, this is a cutoff sample, but generally because of compromises due to the fact that multiple attributes are of interest (more than one question on a survey form), we have quasi-cutoff sampling. (See definitions in Knaub(2014).) These cases, plus “balanced sampling” are considered in Knaub(2013b). “Representativeness” is achieved through the use of a regressor (or regressors), and stratification, or here, careful use of estimation groups (EGs). (See the last five pages of Knaub(2007a) which were added in 2014.) Here, in figures below, we plot points from several PGs on a scatterplot for a given EG. These points on a regression line, even when only a very few of the largest (x-value) cases are available per PG, may appear as one quasi-cutoff sample, depending upon where x-values for any missing y-value data may fall.

Scatterplots can be used for editing/analyses, and to relate that to the sample itself, plotted on the graph, which is also indicative of the regression model to be used for estimation of missing data. Variance is also indicated on such a scatterplot graph. Thus a great deal of information is integrated and immediately visually available on such a scatterplot. One scatterplot here, Figure 2 in section 3, even illustrates seasonality by plotting y-values for two different periods (here, sample months) on the same graph, against the same regressor data (here, annual census data for the same data element).

Ahmed and Kirkendall(1981), and “Option # 2” in Knaub(2011), provide the basis for this simplified small area estimation (SAE). In section 5 we also make comparison to the case of not using SAE, which is “Option # 1” in Knaub(2011a). Here there is some reference to varying degrees of heteroscedasticity (Knaub(2007b), Knaub(2011b)), but the somewhat understated, and perhaps more robust level of heteroscedasticity (Knaub(1993, 2001)) inherent in the model-based classical ratio estimator (CRE), found in Knaub(2005), is the focus here. See Holmberg and Swensson(2001) for more on understating heteroscedasticity.

2. Small Area Estimation for Nonresponse when \( n = N = 1 \)

For Use When Aggregating to a Higher Level

First, re Figure 1, we will consider the case where a category (PG) for which we will want to publish an estimated total, consists of a sample datum from only one possible respondent. The only source of error for this ‘estimate’ is nonsampling error (which is not to be taken lightly). In this case, the sample size, \( n \), and the (sub)population size, \( N \), are both equal to 1 for the State of Vermont (VT) for a given monthly sample collection (not a census for all PGs), with regressor data from an annual census. In the example shown, if the respondent were to nonrespond for Vermont, or respond with a number that did not pass edits and the respondent did not provide data deemed acceptable to replace it, then the classical ratio estimator would be applied as shown in the results of a spreadsheet presented in Figure 1. The estimated RSE in this case was less than 5%, but any bias – the difference between the unknown slope, b-value, that should be used for Vermont, and the b-value for the EG used here – would be unknown. Further, many data users, for purposes of decisions they want to make, inherently assume RSE values much less than 1%. Sometimes even trying to obtain such a low RSE estimate will only drive up nonsampling error. This can actually make RSE estimates start to become larger, but they will go to zero if a census is obtained, no matter how much nonsampling error is then present. For this particular case, however, the SAE estimated RSE is about 4.6%, and the bias is not measureable as no data from the category to be published was actually collected in the monthly ‘sample.’ Thus publication of the SAE
CRE ‘predicted’ value for Vermont is not recommended. But the predicted value would be included in a US-aggregate level total, as well as adding the variance to the US-level variance. (The various geographic regions by EG are independent, so variances are additive.) In the example, Figure 1 below, a test is conducted where we actually have this y-value for Vermont, but we ‘pretend’ not to have it.

**Figure 1**

Small Area Estimation for Nonresponse when \( n = N = 1 \)
For Use When Aggregating to Higher Level

Spreadsheet Test Results:

Four New England States with \( n < 4 \) Each, for Natural Gas Deliveries, in a Given Economic End-Use Sector, For a Given Month, with Regressor Data from an Annual Census

<table>
<thead>
<tr>
<th>Sum_n_all of x</th>
<th>Sum_n_all of y</th>
<th>Sume02 =</th>
<th>MSEw =</th>
<th>sigma2 =</th>
<th>V(b') =</th>
<th>se(b') =</th>
<th>SumX_N-n =</th>
</tr>
</thead>
<tbody>
<tr>
<td>23792084</td>
<td>4004178</td>
<td>787.56627</td>
<td></td>
<td>157.513254</td>
<td></td>
<td>6.6204E-06</td>
<td>3011635</td>
</tr>
</tbody>
</table>

\[ b' = 0.16829875 \]

\[ V(b') = 6.6204E-06 \]

\[ se(b') = 0.00257301 \]

\[ VT \rightarrow Vh*(I^*\cdot T) = 534419149 \]

\[ VT \rightarrow RSE^*in\% = 4.56097592 \]

\[ VT = 3011635 \]

\[ y = 506854.4063 \]

\[ VT_T_SAE = 506,854 \]

[Graph showing data points for Vermont and other New England States]
Vermont (VT) has \( n = N = 1 \). Here we 'pretend' VT nonresponded in the given month. We impute 507K, in whatever the units are here, with an estimated relative standard error (RSE) of approximately 4.6%. The confidential number actually collected differed from this by less than 1.5 RSEs. The bias contribution can be studied historically. (See Knaub(2001).) The contribution to the estimated US level total and relative standard error are taken from \((b')_*(x)\) and \(VL*(T^*_T)\) above, with 'add-ons' to the total from the other censused States.

Note that a prime mark, ‘’,’ or ‘_all,’ either one here, means to include all “borrowed” data. Thus, \( n_{\text{all}} \) means all six y-value observations (one “hidden” under another), when we remove the response from Vermont. These six y-values, along with the corresponding x-values, are used in the calculation of \( b' \).

Note that sometimes it may be convenient to write \( n_{\text{all}} \) for \( n_{\text{all}} \) in equations in this article.

If there had been other data \((x,y)\) pairs for this data element in Vermont, or the other States in this EG, they would have been counted and used as well.

3. **Seasonality**

Here, in Figure 2, we consider data from two monthly samples, six months apart, graphed against the same annual regressor data in one scatterplot, which demonstrates how well some data may behave together when placed in the same estimation group (EG). Knaub(2010) warns that for one regressor, different subsets of data should not be seen to need separate regression lines (b-values), but here those two b-values are for different data collection periods, so this only demonstrates that seasonality far outweighs any perceived difference in States (PGs).
Figure 2

Seasonality Far More Important in this Figure than Differences Between State-level Data

Showing seasonality by graphing data for a given data element, for a month from winter and a month from summer, on the same plot, against the same regressor data (x), from an annual census:

“Close-up” of part above, nearer to origin:
4. Scatterplot Editing

Here, in Figure 3, we consider an example of the use of these scatterplots when editing data.

**Figure 3**

**Scatterplot Editing with Small Area Estimation for Nonresponse - Very Small n**

Some scatterplot graphs with small $n$ and $N$, also for real data, line up tightly on a regression line, like this:

![Scatterplot](image1)

But then, data from the next month, using the same regressor data, may look like these real data did:

![Scatterplot](image2)
Note that weighted least square regression confidence bands about the \( y \)-values, with regression through the origin, as in Knaub(2009), section 4, pages 7 through 9, and shown in figures in Knaub(2012a and 2012b), may be useful here.

By itself, the gold-colored “x-shape-labeled” \((x,y)\) data-point above may not have raised any flags, but here, graphed with these other data, it appears that there is possibly substantial nonsampling/measurement error. Often, establishment survey data seem to have disproportionately large measurement error near the origin (i.e., relatively large measurement error in \( y \), with smaller \( x \), the size measure). Also, the variance for \( y \) about the regression line through the origin should be greater for larger \( x \) (or multiple regressor size measure, say a linear combination). However, here the suspected presence of an outlier for this large \( x \)-value case seems justified, as the estimated residual seems quite large. A reasonable confidence band about the \( y \)-values, calculated as described in Knaub(2009), section 4, should leave the gold “x” point above, out-of-bounds, and in need of confirmation.

Sometimes the data are just generally more variable, as illustrated in the last three pages of Knaub(2011), which is a new section added in 2014. Also, as just noted, often data near the origin have disproportionately large nonsampling error. Including some of them in the sample sometimes causes the estimated relative standard error (RSE) to rise, but inevitably the RSE estimate drops to zero for a census, no matter how large the nonsampling error may be.

**When a State has just one respondent \((n = 1, \text{or maybe even } n=N=1)\), this (such a scatterplot) would be a way to see if the one response obtained appears to fit the market trends.**

5. **Comparing Results With and Without Small Area Estimation**

To use the projected variance tool from Knaub(2012b, 2013a, 2013b) with the simple small area estimation (sSAE) here, we must compare Options # 1 and 2, pages 10 and 11, Knaub(2011a), where Option # 2 is actually the simple, SAE formulation, also inherent in Ahmed and Kirkendall(1981). We would like to see the relative efficiency of using this small area estimation. That will be considered in this section.

Below, either “\_all” or a prime symbol (’’) will be used to indicate places where the “borrowed strength” of the ‘extended’ sample, within one estimation group (EG), will be used. (For \( b’ \), the associated extra-publication group (extra-PG) regressor data are also used.) These data are outside of the subpopulation that are part of a total to be estimated and published, a PG.

For “EG” and “PG,” see Knaub(1999). A flexible, but more complex SAE method is used there, which includes multiple regression, and PGs that are not all completely contained within a given EG. (In that more complex method, PGs can be re-designed and re-published easily, on any occasion.)

Here we consider a one-regressor, ratio (zero intercept) model:

\[
y_i = bx_i + e_{0i}x_i^y
\]

And in particular, the classical ratio estimator (model-based), see Lohr(2010), Chapter 4, where we use \( y = \frac{1}{2} \).
For the variance of the error in predicting the total, $T$, we will designate the estimate of $T$ by $T'$. This is in accordance with the notation used by Maddala(2001), where he uses a “hat” to designate an OLS estimate, but for weighted least squares (WLS) regression, he uses an asterisk. Also, from page 85 of Maddala(2001), and on page 7 of Knaub(2009), we have this for the variance of the ‘prediction’ error for one regressor:

$$V^*_L(y^*_i - y_i) = \frac{\sigma^*_{e_0}^2}{w_i} + x_i^2 V^*(b), \quad \text{where} \quad \sigma^*_{e_0}^2 = \sum_{i=1}^{n} e_{0i}^2 / (n - 1),$$

and when $y_i = bx_i + e_{0i} x_i^y$, we have $w_i = x_i^{-2y}$, $b = \frac{\sum_{i=1}^{n} x_i w_i y_i}{\sum_{i=1}^{n} w_i x_i^2} = \frac{\sum_{i=1}^{n} x_i^{1-2y} y_i}{\sum_{i=1}^{n} x_i^{2-2y}},$

and $V^*(b) = \sigma^*_{e_0}^2 / \sum_{i=1}^{n} x_i^{2-2y}$.

$V^*_L(y^*_i - y_i)$, the variance of the prediction error, is the square of “STDI,” in SAS PROC REG, by the way. For $b$, see the bottom of page 2 in Knaub(2009).

The variance of the prediction error is the basis for $V^*_L(T' - T)$, the variance of the error in “predicting” the total for a PG (or EG, if you wish). Royall and Cumberland(1981) tried other estimators for the variance, but robustness there may be questionable (see Knaub(1992), Figure 1, page 879), as well as less clear to analyze. For better robustness, an underestimate of $\gamma$ may help. (See Knaub(2005, 2011b), and Holmberg and Swensson(2001), regarding the coefficient of heteroscedasticity, $\gamma$, and Knaub(2010) regarding robustness in general. “Representativeness” is discussed in a five page addition to Knaub(2007).)

Note from Knaub(1999), $V^*_L(T' - T)$ is not simply the sum of $V^*_L(y^*_i - y_i)$ over the missing data. The difference is a square of sums, rather than a sum of squares. See Knaub(1999), pages 4 and 5. Consider the following:

$$T = \sum_{i=1}^{n} y_i + \beta \sum_{i=n+1}^{N} x_i + \sum_{i=n+1}^{N} \epsilon_{0i} x_i^y,$$

and $T' = \sum_{i=1}^{n} y_i + b \sum_{i=n+1}^{N} x_i$,

so $T' - T = \sum_{i=n+1}^{N} (b - \beta) x_i + \sum_{i=n+1}^{N} (-\epsilon_{0i} x_i^y)$.

$$V_L(T' - T) = V[\sum_{i=n+1}^{N} (b - \beta) x_i] + V[\sum_{i=n+1}^{N} (-\epsilon_{0i} x_i^y)]$$

$$V[\sum_{i=n+1}^{N} (b - \beta) x_i] = V[(b - \beta) \sum_{i=n+1}^{N} x_i] = (\sum_{i=n+1}^{N} x_i)^2 V(b - \beta)$$

$$V[\sum_{i=n+1}^{N} (-\epsilon_{0i} x_i^y)] = V[\sum_{i=n+1}^{N} (\epsilon_{0i} x_i^y)] = \sum_{i=n+1}^{N} \sigma^2_{e_{0i} x_i^y} = \sigma^2_{e_{0} x_i^2y} \sum_{i=n+1}^{N} x_i^{2y},$$

where this time we do not have a power of a sum, because the factor $\epsilon_{0i}$, unlike $V(b - \beta)$, is subscripted as varying with each value of “$i$,” and could not be moved outside of the summation.

So, $V^*_L(T' - T) = (\sum_{i=n+1}^{N} x_i)^2 V^*(b) + \sigma^*_{e_{0}}^2 \sum_{i=n+1}^{N} x_i^{2y}$, or
\[ V^*_i(T^* - T) = \sigma^{i^2}_{e_0} \sum_{i=n+1}^{N} x^2_i + V^*(b) \left( \sum_{i=n+1}^{N} x_i \right)^2 \]

For \( \gamma = \gamma_z \), using pages 10 and 11 of Knaub(2011a), consider this:

For Option # 1:

\[ T^* = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} b x_i \]

\[ V^*_i(T^* - T) = \sigma^{i^2}_{e_0} \sum_{i=n+1}^{N} x_i + V^*(b) \left( \sum_{i=n+1}^{N} x_i \right)^2 \]

\[ b = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i \]

\[ \sigma^{i^2}_{e_0} = \sum_{i=1}^{n} e_{0i}^2 / (n - 1) \]

\[ V^*(b) = \sigma^{i^2}_{e_0} \sum_{i=n+1}^{N} x_i \]

But for Option # 2:

\[ T'^* = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} b' x_i \]

\[ V'^*_i(T'^* - T) = \sigma^{i^2}_{e_0} \sum_{i=n+1}^{N} x_i + V'^*(b') \left( \sum_{i=n+1}^{N} x_i \right)^2 \]

\[ b' = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i \]

\[ \sigma^{i^2}_{e_0} = \sum_{i=1}^{n} e_{0i}^2 / (n - 1) \]

\[ V'^*(b') = \sigma^{i^2}_{e_0} \sum_{i=n+1}^{N} x_i \]

More generally, for \( y_i = bx_i + e_0 x_i^\gamma \), we have
with \( V^*(b) = \sigma^*_{e_0} \sum_{i=1}^{n} x_i^{2} \) from Knaub (2009), bottom of page 19, and Maddala (1977), pages 259 and 260, and partially from above, and for the SAE here,

\[
V^*_L(T^* - T) = \sigma^*_{e_0} \sum_{i=n+1}^{N} x_i^{2Y} + V^*(b) \left( \sum_{i=n+1}^{N} x_i \right)^2
\]

Using

\[
V^*_L(T^* - T) = \sigma^*_{e_0} \sum_{i=n+1}^{N} x_i^{2Y} + V^*(b) \left( \sum_{i=n+1}^{N} x_i \right)^2
\]

and substituting for \( V^*(b) \),

\[
V^*_L(T^* - T) = \sigma^*_{e_0} \sum_{i=n+1}^{N} x_i^{2Y} + \frac{\sigma^*_{e_0} \sum_{i=1}^{n} x_i^{2}}{\sum_{i=1}^{n} x_i^{2-2Y}} \left( \sum_{i=n+1}^{N} x_i \right)^2
\]

we then obtain

\[
V^*_L(T^* - T) = \sigma^*_{e_0} \left[ \sum_{i=n+1}^{N} x_i^{2Y} + \frac{\sum_{i=n+1}^{N} x_i^{2}}{\sum_{i=1}^{n} x_i^{2-2Y}} \right]
\]

For SAE, redefining \( \sigma^*_{e_0} \) as \( \sigma^*_{e_0} \), where

\[
\sigma^*_{e_0} = \sum_{i=1}^{n_{all}} e_{0i}^2 / (n_{all} - 1)
\]

\[
V^*_L(T^* - T) = \sigma^*_{e_0} \sum_{i=n+1}^{N} x_i^{2Y} + V^*(b') \left( \sum_{i=n+1}^{N} x_i \right)^2
\]

\[
V^*_L(T^* - T) = \sigma^*_{e_0} \left[ \sum_{i=n+1}^{N} x_i^{2Y} + \frac{\sum_{i=n+1}^{N} x_i^{2}}{\sum_{i=1}^{n_{all}} x_i^{2-2Y}} \right]
\]
Then (borrowing from the somewhat analogous idea of Kish’s “design effect”), we have a small area effectiveness measure $saeff$, where

$$saeff = \frac{V_L'(T^* - T)}{V_L'(T^* - T)} = \frac{\sigma_{e_0}^2}{\sigma_{e_0}^2} \left( \frac{\sum_{i=1}^{N} x_i^2 (\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{all} x_i})^2}{\sum_{i=1}^{N} x_i^2} + \frac{(\sum_{i=1}^{n} x_i)^2}{\sum_{i=1}^{all} x_i^2 - 2\sum_{i=1}^{n} x_i}\right) = \frac{\sigma_{e_0}^2}{\sigma_{e_0}^2} \left( \frac{\sum_{i=1}^{N} x_i^2 (\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{all} x_i})^2}{\sum_{i=1}^{N} x_i^2} \right)$$

(Note definitions in Kish(1995), page 56, for “Definitions of Design Effects,” to see the similarity.)

For $Y = \gamma_x$ the model-based classical ratio estimator (CRE), and letting $T_x = \sum_{i=1}^{N} x_i$, the sum of the regressor data for the population of interest:

$$saeff_{CRE} = \frac{\sigma_{e_0}^2}{\sigma_{e_0}^2} \left( \frac{\sum_{i=1}^{N} x_i^2 (\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{all} x_i})^2}{\sum_{i=1}^{N} x_i^2} \right) = \frac{\sigma_{e_0}^2}{\sigma_{e_0}^2} \left( \frac{1 + (T_x - \sum_{i=1}^{n} x_i) / \sum_{i=1}^{n} x_i}{1 + (\sum_{i=1}^{all} x_i - \sum_{i=1}^{n} x_i) / \sum_{i=1}^{all} x_i} \right) = \frac{\sigma_{e_0}^2}{\sigma_{e_0}^2} \left( \frac{\sum_{i=1}^{n} x_i}{T_x / \sum_{i=1}^{n} x_i} \right) \left( \frac{T_x - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{all} x_i} \right)$$

Here we will let $MSE_W = \sigma_{e_0}^2$. See Kutner, Nachtsheim, and Neter(2004) for the use of the subscript “$W$.” Note that SAS PROC REG, when used for weighted least squares regression, here the classical ratio estimator with $w = 1/x$, provides this estimate labelled as just $MSE$. However, it really is $\sigma_{e_0}^2$.

So ...

$$saeff_{CRE} = \frac{MSE_W}{MSE_W} \left( \frac{\sum_{i=1}^{n} x_i}{T_x / \sum_{i=1}^{n} x_i} \right) \left( 1 + \frac{T_x - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{all} x_i} \right)$$

Therefore

$$\sqrt{saeff_{CRE}} = \sqrt{\frac{MSE_W}{MSE_W} \left( \frac{\sum_{i=1}^{n} x_i}{T_x / \sum_{i=1}^{n} x_i} \right) \left( 1 + \frac{T_x - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{all} x_i} \right)}$$

is the fraction of the standard error that remains when using small area estimation (SAE). Hopefully it is less than 1, as there is still bias to consider. However, the $MSE_W$ here does indirectly consider bias, to an extent, and the sum of the estimated residuals does not equal zero for the SAE. For the usual CRE, it does. See Knaub(2005).

Further, these standard errors are not designed to account for nonsampling error, but do to some extent, that extent diminishing as the sample size $n$ approaches the population size $N$. This is because,
just as in design-based estimation, where the finite population correction factor is used, variance from the sample is applied to the part of the population out-of-sample, to estimate the error that is due to the fact that we did sample (i.e., sampling error). Variance is influenced by nonsampling error more fully before the “correction.” In model-based estimation, the summations over N-n represent making this “correction.” (See Knaub(2008).)

Cochran(1977), page 158, citing Brewer(1963) and Royall(1970), notes that here we are considering variance with a “model-unbiased” estimator. The bias we consider here is the degree to which the data do not fit the model. Shmueli(2009), at one time found on the Internet, gave a good explanation for this. Bias from using small area estimation, when certain data sets are best modeled separately but here are put together under the same model, can be seen to impact the estimated variance of the error in the ‘predicted’ total directly.

When the simple model here is written as

\[ y_i = bx_i + e_{0i}x_i^Y \]

and

\[ T^* = \sum_{i=1}^{n} y_i + b \sum_{i=n+1}^{N} x_i \]

then we may write the estimated variance of the error in the estimated total again as

\[ V^*_E(T^* - T) = V^*(b) \left( \sum_{i=n+1}^{N} x_i \right)^2 + \sigma^*_e^2 \sum_{i=n+1}^{N} x_i^{2Y} \]

where the estimated variance for the error in predicting the total, using the model \( y_i = bx_i + e_{0i}x_i^Y \), can be considered in two parts, with \( V^*(b) \left( \sum_{i=n+1}^{N} x_i \right)^2 \) corresponding to the \( bx_i \), and \( \sigma^*_e^2 \sum_{i=n+1}^{N} x_i^{2Y} \) corresponding to the \( e_{0i}x_i^Y \). Thus, it is important to compare \( V^*(b) \) to \( V^*'(b) \), and \( \sigma^*_e^2 \) to \( \sigma^*_{e2} \).

Using primes to designate the above SAE related variables, and \( n_{all} \) to designate the size of the sample from the population of interest, plus the “borrowed” sample, as applied to small area estimation, and \( \sqrt{saeff} \) to represent the small area estimation efficiency by providing the fraction of the standard error that occurs using small area estimation as opposed to not borrowing strength, we obtained
\[
\sqrt{\text{saef}_{\text{CRE}}} = \sqrt{\left(\frac{\text{MSE}_W'}{\text{MSE}_W}\right) \left(\frac{\sum_{i=1}^{n} X_i}{T_x} \right) \left(1 + \frac{T_x - \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n_{all}} X_i}\right)}
\]

where \( T_x = \sum_{i=1}^{N} x_i \).

To estimate this, we need \( \text{MSE}_W, \text{MSE}_W', \sum_{i=1}^{N} x_i, \sum_{i=1}^{n} x_i, \sum_{i=1}^{n_{all}} x_i \).

However, it may be easier to recognize the influence of the various terms by simply viewing it this way:

\[
\text{saef}_{\text{CRE}} = \left(\frac{\text{MSE}_W'}{\text{MSE}_W}\right) \left(1 + \frac{(T_x - \sum_{i=1}^{n} x_i)/\sum_{i=1}^{n_{all}} X_i}{T_x - \sum_{i=1}^{n} x_i/\sum_{i=1}^{n} X_i}\right)
\]

or

\[
\text{Eq 1:}
\]

\[
\text{saef}_{\text{CRE}} = \frac{V_{L}'' (T'' - T)}{V_{L} (T - T)} = \left(\frac{\text{MSE}_W'}{\text{MSE}_W}\right) \left(1 + \frac{(\sum_{i=n+1}^{N} x_i)/\sum_{i=1}^{n_{all}} X_i}{(\sum_{i=n+1}^{N} x_i)/\sum_{i=1}^{n} X_i}\right)
\]

The expression on the right in the right-most parentheses will definitely be less than unity, if ‘strength is borrowed.’ The rest depends on the variability of the random factors of the estimated residuals. Note Figure 4. The \( \text{MSE}_W, \) and \( \text{MSE}_W' \) values are consistent with the above equation.

Consider a case, which in an extreme situation might look like Figure 2, except say there were only one sample, where the two obvious regression lines represented different regions, instead of different samples. If those data were treated together, with only one regression line through them, the resulting bias would inflate the \( \text{MSE}_W' \) estimate.
Figure 4

SAE vs no SAE

Using vs not using small area estimation (SAE)

Consider this scatterplot again:

Suppose these were all the data for this given estimation group, except that there were a fourth out-of-sample case for State A, where the x-value was approximately 340,000. (Respondent level data are not given precisely here, because of confidentiality arrangements for the y-values.)

Considering only the **State A** data yields the following, from the dataset used to make the above scatterplot graph:

\[ T^* = 6,509,834 \text{ or approximately 6510 K units} \]
\[ b = 0.1586 \]
\[ \text{s.e.}(b) = 0.0059 \]
\[ \text{RSE}^* = 0.34\% \]
\[ MSE_W = 1433 \]

Considering all data \((n_{all} = 10, N_{all} = 11)\) in estimating the total for **State A** \((n = 3, N = 4) – SAE – \) yields the following:

\[ T' = 6,511,001 \text{ or approximately 6511 K units} \]
\[ b' = 0.1620 \]
\[ \text{s.e.}(b') = 0.0029 \]
\[ \text{RSE}'^* = 0.21\% \]
\[ MSE_W' = 568 \]
From Eq 1, \( s_{\text{CRE}} = \frac{\text{MSE}_{w}}{\text{MSE}_{w}'} \left( \frac{1}{1 + \frac{Q_N}{\sum_{i=1}^{n+1} X_i}} \right)^{\frac{1}{2}} \). Because there is only one, relatively small out-of-sample case, \( s_{\text{CRE}} \) should be just slightly smaller than \( \frac{\text{MSE}_{w}'}{\text{MSE}_{w}} \). So \( \left( \frac{0.21}{0.34} \right)^2 \) should be slightly smaller than \( \frac{568}{1433} \), as \( T' \) is nearly equal to \( T^* \).

Because \( \left( \frac{0.21}{0.34} \right)^2 \approx 0.3815 \) and \( \frac{568}{1433} \approx 0.3964 \), this indicates these results are correct.

Also note that the difference in estimated total is only \( \frac{6,509,834 - 6,511,001}{6,509,834} = 0.018\% \). However, the coverage here was over 99\%. Still, the RSE estimates seem a bit large. This appears consistent with the author’s \textit{ad hoc} observation over years that these models tend to overestimate error for very small error cases.

Recall also that the characteristic of classical ratio estimation discussed in Knaub(2005), and noted elsewhere, whereby the sum of the estimated residuals is always exactly equal to zero, would not hold for just one PG under an EG model here.

**Summary:**

In this paper we explore some aspects of the connection between (1) the sample selected, (2) model-based estimation – notably the classical ratio estimator, and (3) data editing/analyses. This is done here in the context of small area estimation. Accordingly, figures show a special nonresponse case, seasonality, scatterplot editing, and a basis for the comparison of estimations with and without small area estimation methodology. From each of these figures we can perceive (1) how the sample is laid out, (2) how those points contribute to estimation, and (3) visually analyze aspects of the data.

**Acknowledgements:** The author is grateful for the pleasure of working with a number of creative, insightful, and inspiring people. Work with Joel Douglas at the US Energy Information Administration (EIA) (transferring to the Federal Energy Regulatory Commission), such as that done on variance estimation, and especially with what we have called “quasi-cutoff sampling,” has been “awesome” for me. (I still think you came up with that name, Joel.) Association with Samson Adeshiyian, Carrie Hughes-Cromwick, Jason Worrall, and others, including energy experts such as Jennifer Wade, has made the work leading to this very enjoyable. Nancy Kirkendall and others were helpful in the references below. And as usual, I want to thank my very kind mentor, Ken Brewer, for all he has done over a number of years.
References:


- Short version:

- Short version:


Note that InterStat articles by this author are being copied to ResearchGate.