Comparison of Model-Based to Design-Based Ratio Estimators

James R. Knaub, Jr.
Retired

Abstract
Ratio estimation is often useful for Official Statistics regarding energy, and for agriculture, econometrics, and perhaps many other applications in business, social science, and other areas. Notably, ratio estimation is very often useful for highly skewed establishment survey populations where, per Brewer(2002), mid-page 111, there should be at least as much implicit heteroscedasticity as for that of the classical ratio estimator (CRE). The concepts of design-based and model-based ratio estimation and sampling are reviewed and compared. Meaningfulness might be enhanced by understanding this comparison. Note that here the design-based case is actually model-assisted, but is being contrasted with the strictly model-based methodology, where probability of selection does not enter into the estimation, actually 'prediction,' of totals, and may or may not be used for sample selection.[1] These model-based and design-based interpretations of the CRE, their corresponding concepts of variance and bias, with relation to sampling and estimation, are reviewed, and extensions of these estimators are also considered. Simple random sampling, cutoff, and unequal probability of selection methodologies are of interest. Stratification is often highly useful with either approach. Even if a regression model is not explicitly considered, this review considers the role it still plays. The relationship of heteroscedasticity, explicitly addressed in model-based estimation, to cluster sampling for unequal-sized censused clusters, is a point of interest: Each observation in a model-based sample may be treated as a cluster unit for which we have a census. (Again, see Brewer(2002), mid-page 111.)

Key Words: unconditional distribution, conditional distribution, expansion factor, regression coefficient, optimal sampling, multipurpose surveys, bias, variance, coefficient of heteroscedasticity

1. Applications and Introduction

The concepts of probability sampling (design-based) and estimation, and prediction-based sampling and estimation, can be found in various subject matter areas. In soil science, articles such as Rossel, et.al.(2016) use the model-based, model-assisted, and design-based concepts for estimation/prediction with a probability sample, and numerous other papers in soil science can be found, such as those whose authors include D.J. Brus and/or J.J. de Gruijter. Note also that model-based, model-assisted, and design-based concepts are also found in forestry. See Warren(2004). But what about simply a ratio estimator with either probability-based or model-based sampling?

[1] For strictly model-based sampling, the y_i are selected by the values of the x_i (see bottom of page 158 in Cochran(1977)), but randomized sampling, or better, "balanced sampling," may be used to reduce bias, at the expense of an often extreme increase in variance for skewed data.

The ratio estimator, considered either (1) with a design-based approach to estimation and sampling (thus model-assisted), or (2) with strictly model-based sampling and estimation, or (3) with perhaps probability sampling and model-based estimation, that is, prediction, which, re calculations, could only matter for variance estimation, is useful in various survey statistics and related applications. For any finite population, if data are available for the population for one (or more) auxiliary/regressor variable(s), which correlates to the data for which a sample is to be collected, and the relationship is linear and includes the origin, then this may be used to advantage to obtain more efficient estimates of totals or means. Fortunately, this is often the case, or nearly the case, particularly for Official Statistics.

No matter how data are collected, or considered at the time with regard to design, one may always use model-based variance estimates[2] and straight forwardly tie different strata together under the model-based approach.

2. Overall Concepts

Even design-based ratio estimation is model-assisted. There is a substantial advantage to having auxiliary data to aid a randomized sampling and estimation method, as one will never know if an influential datum, or even portion of the population, may have been missed by a sample, which would greatly impact on estimation and estimation of accuracy. A variance estimate for an estimated total, for example, is based on the variance found in the sample, which is then projected upon the finite population. If influential data are missed, or overrepresented, this is a problem. Auxiliary or regressor data on the population may be very helpful. In particular, unusual members of the population may be identified and properly considered. Note that the smaller the sample, the more likely this could be particularly important. Brewer(2014) notes that smaller samples are more likely to benefit from models. But even with a design-based approach, adjusting the estimate by using a ratio means using data on the entire population, As long as the auxiliary data, x, have some connection to the variable of interest, y, we are not so likely to miss the influence of an individual member or set of members of the population, ignorance of which may greatly degrade the performance of unaided random sampling.

Once you find that you have good auxiliary data, which could be called regressor (independent variable/x/predictor) data, one needs to decide, "Do I want to assume that a model applies, or do I want to assume that data selected at random "represent" other data? KRW Brewer and others combine these approaches. For an entertaining overview and historical perspective, see Brewer(2014). (Notice there that the only model Ken Brewer specifically illustrates is the CRE.)

Because any ratio estimate is at least model-assisted, the devastation that could occur with missing what might be described as anomalous members just considered above, is greatly mitigated, and the decision will become one of a tradeoff of bias for variance reduction, with a smaller sample possible with the model-based approach, *i.e.,* prediction.

For design-based ratio estimation, *i.e.*, a randomized selection of y_i and correlated x_i, we make use of sampling $y_i - \hat{R}x_i = e_i$. Both $\boldsymbol{y_i}$ and $\boldsymbol{x_i}$ are random variables, but only in the

[2] See the Greek population study in Deming(1943/1964), Chapter 12, alluded to in Cochran(1977), on page 160.

sense of sample selection, as considered in Sukhatme(1954), pages 139-140. In Thompson(2012), on pages 94, 105, respectively, he notes the fixed nature of x and y for design-based sampling. These e_i are generally not optimally distributed for simple random sampling (SRS). SRS corresponds to homoscedasticity, whereas R, Sukhatme(1954), page 139, corresponds to the CRE below. For design-based sampling, the estimator is shown to be biased for simple random sampling (SRS) when not represented by a line through the origin. (See Sukhatme(1954), page 143). This is also the case with a generalized ratio estimator, one which accommodates unequal probabilities of selection, as noted in Thompson(2012), page 102, where he notes that any such ratio estimator is not design-unbiased, for the same reason that Sukhatme(1954) noted this for SRS: division of one design-unbiased estimator by another is not unbiased.

Optimal designs consider that $|y_i - \hat{R}x_i|$ generally increases with increasing x_i. That is, there is heteroscedasticity which is quite prevalent for skewed data from establishment surveys, where ratio estimation is often helpful, notably when auxiliary data are available from, say, an annual census, and the sample is for a monthly data collection on the same data elements. Thus an optimal design-based sample would involve unequal probability sampling, based on knowledge of heteroscedasticity, as considered below, in this paper.

For model-based ratio estimation, for a data group/stratum where one model applies, y_i may be sampled in any manner. Here we need $y_i - bx_i = e_i = e_{0_i} w_i^{-0.5}$, a regression. In such a case, we consider that y is a random variable, but x is not. The $\boldsymbol{y_i}$ are conditional on given $\boldsymbol{x_i}$. This assumes a line through the origin, as in the unbiased case for a design-based ratio estimator. In Sukhatme(1954), pages 143-144, this condition for bias to 'vanish' for SRS makes sense because weighted least squares (WLS) regression, and ordinary (homoscedastic) least squares (OLS) regression are both unbiased for b. That is, we see on pages 138 to 143 of Sukhatme(1954), a derivation of the design-based ratio estimator which shows it is unbiased when we have a linear regression through the origin with the regression coefficient being homoscedastic. But then on page 144 he argues that one can use an estimated regression coefficient, shown as equation (1) on page 139, which is clearly heteroscedastic, as it is the ratio (regression coefficient) for the CRE. However, in Maddala(2001), on pages 207 to 208, we see that when the WLS ratio estimator is required, the OLS ratio estimator is less efficient, but it is still unbiased. Thus the results shown on pages 143 and 144 in Sukhatme(1954), from the development of pages 138 to 143, leave us with an unbiased regression coefficient when we have linear regression with a zero intercept.

Note that for the model-based classical ratio estimator (CRE), which is really regression prediction where the regression weight may be described as $w_i = x_i^{-2\gamma} = x_i^{-1}$, this corresponds to the design-based CRE. Perhaps in the model-based case we should say "Classical Ratio Prediction (CRP)."

To apply a model-based CRE, the y_i are sampled, but not necessarily randomly. Stratification may be particularly important, however, to be certain that each model is only applied to the part of the population relevant to that model. (This model-relevance is important to both stratification and "borrowing strength" for small area estimation, as the latter is explained in Knaub(1999), for examples involving estimations for (sub)totals of hydroelectric generation. Note there that a special variance estimate was used for purposes of flexibility, unrelated to this paper, but the overall concept still applies.)

The x_i for $i = 1, N$ are generally considered to be "known." However, if errors-in-variables are considered for both y_i and x_i in model-based ratio estimation, b is biased downward, as shown in Maddala(2001), pages 438 to 440, Section 11.2, "The Classical Solution for a Single-Equation Model with One Explanatory Variable." If the x_i are uncertain from an inverse regression perspective, then from Deming(1943/1964), for $w_i^{-0.5}$, he uses $W_i^{-0.5} = \left(b^2 w_{x_i}^{-1} + w_{y_i}^{-1}\right)^{0.5}$, where for the CRE, $w_{y_i} = x_i^{-2\gamma} = x_i^{-1}$. Finally, if the x_i are generally observed, but some values might be predicted from a previous process, as may often be the case for official statistics, Joel Robert Douglas suggested directly considering the increase in uncertainty that would result. (Douglas and Knaub(2010).)

Joel Douglas also created a "tiered" system of estimation for the US Energy Information Administration (EIA), where only the best regressor or regressors was/were used in each ratio estimation across a population, or set of subpopulations. See Douglas(2013), while available, and some slides attributed to Joel Douglas in Knaub and Douglas(2010).

An excellent discussion regarding the model-based CRE is found in Cochran(1977), pages 158 to 160, Section 6.7, "Conditions Under Which the Ratio Estimator is a Best Linear Unbiased Estimator." Note on page 125 of Brewer(2002), equation 8.5 shows us that in general, *b* must be the best linear unbiased estimator (BLUE), for the prediction of y to be from the best linear unbiased predictor (BLUP).

3. Methodology

Basically, for the Classical Ratio Estimator (CRE), we have the following:

Design-based estimator, based on simple random sampling:

$$\widehat{\boldsymbol{Y}}_{\widehat{\boldsymbol{R}}} = \left[\frac{\sum_{i=1}^{N} \boldsymbol{x_i}}{\sum_{i=1}^{n} \boldsymbol{x_i}}\right] \sum_{i=1}^{n} \boldsymbol{y_i} = \hat{R}X$$

Note that $\hat{R}$ is not $\left[\frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} x_i}\right]$. Actually, $\hat{R} = \left[\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\right]$, which is b in the model-based approach.

Some have been known to refer to $\left[\frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} x_i}\right]$ as the "ratio," but that should be reserved for $\hat{R}$ or b. Let us refer to $\left[\frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} x_i}\right]$ as an expansion factor, like $\frac{N}{n}$, but weighted by the size-measures, x_i.

Model-based estimator, actually based on prediction (i.e., regression), for any reasonable[3] sample:

$$\boldsymbol{T}^* = \left[\frac{\sum_{i=1}^{n} \boldsymbol{y_i}}{\sum_{i=1}^{n} \boldsymbol{x_i}}\right] \left(\sum_{i=1}^{N} \boldsymbol{x_i}\right) = bX$$

[3] Just as when determining the stratification for design-based stratified random sampling, one relies on knowledge of auxiliary data to determine what might be "reasonable."

3.1 Design-Based Ratio Estimator:

For simple random sampling (SRS), the usual *expansion factor,* $\frac{N}{n}$, which is the inverse of the sampling fraction, is used for the estimation of the population total as follows:

$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i$

For the ratio estimator, we have $\hat{Y}_{\hat{R}} = \left[\frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} x_i}\right]\sum_{i=1}^{n} y_i$, so $\frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} x_i}$ acts as an expansion factor for the sample sum of y_i, in this paired data sample, to estimate for the population total of the y_i. Such an expansion (or "expanding") factor is part of what has been varyingly presented as a weighting procedure, at least in Kish(1965), pages 203 and 204, and Lohr(2010), p. 122. So instead of referring to $\frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} x_i}$ as a 'ratio,' we may consider it a kind of 'expansion factor,' based on auxiliary data, and *n*, and *N*. Note that on page 86 of Raj(1968), Des Raj states that when we do not 'use' *x* in $\hat{Y} = \sum_{i=1}^{N} x_i \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i$, then we have $\hat{Y} = N\sum_{i=1}^{n} y_i / n$, thus giving $\sum_{i=1}^{N} x_i / \sum_{i=1}^{n} x_i$ a role analogous to N/n.

More generally, $\hat{Y}_{\hat{R}} = \left(\sum_{i=1}^{n} \frac{y_i}{\pi_i}\right)\left(\frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} \frac{x_i}{\pi_i}}\right)$ is the design-based ratio estimator, where $\left(\frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} \frac{x_i}{\pi_i}}\right)$ adjusts the Horvitz-Thompson estimator. This can be written as $\hat{Y}_{\hat{R}} = \left[\frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}}{\sum_{i=1}^{n} \frac{x_i}{\pi_i}}\right]\sum_{i=1}^{N} x_i = \hat{R}X$, and Steven K. Thompson, attributes this to a paper by Ken Brewer in 1963, and later work by others, as noted on page 102 in Thompson(2012).[4]

Also on that page, at the bottom, Thompson states that **the variable which we are considering is not y_i, but $y_i - Rx_i$,** where $R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$. This is an important concept which is often not given enough emphasis, or even left as an implicit truth, which would best be more explicitly recognized.

[Curiously, it seems that both Deming(1950,1966), page 165, and Hansen, Hurwitz, and Madow(1953), pages 107 and 108, reverse the currently accepted order of the use of the variable labels *x* and *y*, when considering a design-based approach.]

Another way to describe a design-based ratio estimator, shown on page 51 of Chambers and Clark(2012), is a weight, w_i there, which has the same expansion factor-like appearance. There, if z is a size measure, then $w_i = \sum_U z_j / \sum_s z_i = \frac{N\bar{z}_U}{n\bar{z}_s}$, where U represents the population (universe), and s is the sample.

Extension: The Chain Ratio-Type Estimator: $\hat{Y}_{cr} = \left[\frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} x_i}\right]^{\alpha}\sum_{i=1}^{n} y_i$, where α is chosen to minimize the mean square error for a given application. See Knaub(2015).

[4] Brewer(1963) is also noted in Cochran(1977), on page 158, for its landmark presentation of model-based ratio estimation/prediction.

3.2 Model-Based Ratio Estimator:

$T^* = \left(\frac{\sum_{i=1}^{n}\frac{y_i}{\pi_i}}{\sum_{i=1}^{n}\frac{x_i}{\pi_i}}\right)\left(\sum_{i=1}^{N} x_i\right) = \hat{R}X$ is design-based, say, model-assisted design-based, but ignoring probability of selection, we have the following:

$T^* = \left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}\right)\left(\sum_{i=1}^{N} x_i\right) = bX$, where T^* is a realization of a random variable, T. (Note Knaub(2013).) This is the model-based classical ratio estimator (CRE), which uses weighted least squares (WLS) regression, with a specific degree of heteroscedasticity, as will now be described:

The derivation for the regression weight in the one-regressor case is shown in a number of places, including the bottom of page 2 in Knaub(2009).

$y_i = bx_i + e_{0_i} w_i^{-0.5}$, is often usefully written as $y_i = bx_i + e_{0_i} x_i^{\gamma}$, and for $\gamma = 0.5$, we have the regression coefficient in the model-based classical ratio estimator (CRE), $b = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$. (See Knaub(2011), regarding the coefficient of heteroscedasticity, γ.)

In general, $b = \frac{\sum_{i=1}^{n} x_i y_i w_i/\pi_i}{\sum_{i=1}^{n} x_i^2 w_i/\pi_i}$, and ignoring sample selection design, $b = \frac{\sum_{i=1}^{n} x_i y_i w_i}{\sum_{i=1}^{n} x_i^2 w_i}$, where we use regression weight $w_i = x_i^{-2\gamma}$. For the CRE, $\gamma = 0.5$, thus resulting in $b = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$. See Knaub(2011) for more on the history of $w_i = x_i^{-2\gamma}$ as a regression weighting system, and note its use in explaining within cluster variance shown, for example, in Cochran(1977), on page 243, with $g = 2\gamma$, which can be traced at least back to Fairfield Smith(1938) in an agricultural setting. For within cluster variance, S_w^2, and M number of elements in the cluster, William Cochran uses $S_w^2 = AM^g$ on page 243. Note that from $y_i = bx_i + e_{0_i} x_i^{\gamma}$ we have that the variance of the estimated residuals is proportional to $x_i^{2\gamma}$, which corresponds to M^g in Cochran. Further, in Murthy(1967), on pages 299 and 300, there is described under cluster sampling, a variance for proportions, which also considers this research in agricultural application(s), also providing references, and with a goal of arriving at the best cluster size.

Note that the celebrated 1948 Greek population study by Jessen and others, including W. Edwards Deming, noted by Cochran(1977) on page 160, is given a chapter in Deming(1950, 1966): Chapter 12, "A Population Sample for Greece," pages 372 through 398. Though, on page 9, Deming states that his book only considers probability-of-selection-based samples, he does devote a good deal of attention to modeling when considering variance, and the Greek population study finds a useful application for $\gamma = 1.0$, such that $b = \frac{\sum_{i=1}^{n} x_i y_i w_i}{\sum_{i=1}^{n} x_i^2 w_i}$, with $w_i = x_i^{-2}$, is helpful. Therefore we may consider $b = \frac{\sum_{i=1}^{n} y_{i/} x_i}{n}$, at least with regard to variance there.

A completely model-based approach considers whether or not there is good partition of data by "estimation group" as in Knaub(1999), where variance is also described, and a

special approximation is included for operational flexibility if needed for a stressed data system.

We should keep in mind that x_i is used as a 'size' measure in the regression weight. If we use multiple regression, we could use predicted-*y*, or some other combination of regressors as the measure of size, instead of *x*. (See Särndal, Swensson, and Wretman(1992), page 232, for more on the use of a linear combination of these regressors, for multiple linear regression.)

The coefficient of heteroscedasticity, γ, for surveys, tends toward $0.5 \leq \gamma \leq 1.0$, with $\gamma = 0.5$ resulting in the CRE. This is analogous to independence of elements within a larger unit (cluster), re Brewer(2002), p. 111. There he explains this by saying that a larger unit in a sample could be considered to be a conglomerate of smaller units. Brewer considers the case of retail stores. If larger units were like a conglomerate of independently controlled retail stores under a given larger retail name, then their variances would be simply additive, which implies $\gamma = 0.5$, where we have $\sigma_i^2 \propto x_i$. Larger values of γ result when there is more central control, as explained by Ken Brewer on that page.

3.3 Calibration:
The ratio estimator is the simplest example of a cosmetically calibrated estimator, i.e., one where the estimator can be "interpreted" from a prediction-based viewpoint. See Brewer(2002), page 104 regarding that. Also, see page 19 in Knaub(2012). For a simple explanation of calibration regarding a ratio estimate, see Lohr(2010), page 132. For calibration to occur, the 'estimate' of $\sum_{i=1}^{N} \boldsymbol{x_i}$ must be exact when we apply the ratio. Note, for example, that in that case, $\left[\frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} x_i}\right] \sum_{i=1}^{n} \boldsymbol{y_i}$ becomes $\left[\frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{n} x_i}\right] \sum_{i=1}^{n} \boldsymbol{x_i} = \sum_{i=1}^{N} \boldsymbol{x_i}$.

3.4 'Optimal' sampling:
This brings us to the consideration of 'optimal' sampling 'design' for both probability-of-selection-based ('design-based'), and regression-prediction (model-based) ratio estimation:

Design-Based: Probability Proportional to $\boldsymbol{x_i^{\gamma}}$.

Model-Based: Cutoff Sampling for largest $\boldsymbol{x_i}$, when appropriate. (But the sample may need to be "balanced" on the mean of x_i, *i.e.*, it may be necessary to require that for the sample, $\sum_{i=1}^{n} \boldsymbol{x_i}/n = \sum_{i=1}^{N} \boldsymbol{x_i}/N$, if bias is a particular problem, say if data falling under different models are treated under one ratio model. Such a problem is shown in Knaub(1999), for example, but solved there by properly identifying subpopulations for model applications. - Balanced sampling can cause a tremendous increase in variance. See Knaub(2013).)

3.4.1 Considering probability of selection based sampling:
First, recall that the variable which we are considering is $y_i - Rx_i$, not y_i. Thus, with unequal probabilities of selection, using the Horvitz-Thompson estimator, one can use variance estimators associated with that.

Now we consider page 254 in Särndal, Swensson, and Wretman(1992), and other parts of that book referenced from that page. For the ratio estimator, R is $\frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$, so for the classical ratio estimator, this is modeled as $y_i = bx_i + e_{0_i} x_i^{\gamma}$, where in the estimated residual and in the formulation for b, to be explained, we have $\gamma = 0.5$. Besides $b = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$, which corresponds to simple random sampling for design-based considerations, we may use $b = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}}{\sum_{i=1}^{n} \frac{x_i}{\pi_i}}$, which is $b = \frac{\sum_{i=1}^{n} x_i y_i w_i / \pi_i}{\sum_{i=1}^{n} x_i^2 w_i / \pi_i}$, with probability of selection π_i, survey weight π_i^{-1}, regression weight $w_i = x_i^{-2\gamma}$, and γ = ½. Optimally efficient sampling for a probability of selection based approach in this situation would not be simple random sampling (SRS). Instead it would be probability of selection proportional to $x^{0.5}$. Särndal, Swensson, and Wretman(1992) calls this $\pi p\sqrt{x}$ (probability proportional to square root of x) sampling. They also note, also on page 254, that for probability proportional to size sampling, we can write πpx. This is because x is a measure of size. Here, we consider the coefficient of heteroscedasticity: γ. The "residuals" (borrowing from model-based language), $y_i - Rx_i$, are estimated by $y_i - bx_i$. But **for the design-based case, we are really looking at probability of selection for these $y_i - bx_i$ values,** and **because these cases are generally heteroscedastic, this is the basis for using unequal probability sampling.** Because these estimated residuals, $e_{0_i} x_i^{\gamma}$, are the sampling unit measures of interest, in general, for optimal design-based sampling, we need probability proportional to x_i^{γ} sampling: denoted as πpx^{γ}. If γ = 1, we have probability proportional to size (PPS) sampling.

3.4.2 Cluster sampling analogy:

One might think of sampling for ratio estimation, either randomized or purposeful/balanced/cutoff, in a skewed population, as if one had a one-stage cluster sample. Each selected unit might be considered to be a cluster with x elements, for which we have a census. This directly links Ken Brewer's explanation (Brewer(2002), page 111) as to why γ should generally not be less than 0.5, with the concept of within cluster variance, as noted, say, in Cochran(1977), on page 243, and discussed above. There Cochran mentions some references and attempts to quantify the within cluster variance, and states that empirical studies in agriculture have shown that the M elements within a cluster can be related to the variance for these elements, within that cluster unit, S_w^2, such that $S_w^2 \propto M^g$, where the g on that page is equal to 2γ here. See Knaub(2011), page 400, and discussion above.

3.4.3 Considering model-based sampling:

For the (strictly) model-based case, it has been shown in numerous publications that the most efficient sampling technique is a cutoff sample of the members of the population of largest size, x. Perhaps the first instance is found in Brewer(1963), and noted in Cochran(1977), on page 160 regarding the classical ratio estimator where he references an estimate of variance he provides on page 159, equation (6.26).[5] The drawback is that ***if*** the model does not apply well to the prediction of y for the smaller x-value members of the "estimation group" (see Knaub(1999) to which a ratio model is applied, then there may be substantial bias. However, this can be overcome with proper 'stratification,' and is

[5] Cochran(1977), equation (6.26) estimates variance as $\frac{\lambda(X - n\bar{x})X}{n\bar{x}}$, which is smaller with larger $n\bar{x}$.

mitigated by the degree to which this can occur when we definitely expect y to be zero if *x* is zero. That is, if we have collected the y-values for the largest x-value members of the population, and we know the origin should be included, this limits how far astray a missing y-value might be. See Knaub(1999) and Knaub(2014a), for example, regarding grouping data for borrowing strength, or separating data by group for stratification, as would be helpful here. Graphics in Knaub(2014b) also show how well data from different areas may or may not fit under a ratio model. Also see the figure on page 9 in Knaub(2010). In Section 5.5 of Chambers and Clark(2012), Figure 5.3 on page 58 shows other examples of using scatterplots to illustrate the application of linear regression, with different estimates of the regression coefficient, *b*, for sugarcane farms, located in four different "growing regions."

Bias may be present for model-based ratio estimation (*i.e.*, prediction) with cutoff, or even quasi-cutoff sampling, but can be quite limited by careful model application. For instance, in Knaub(1999) it was found that totals for hydroelectric generation could be estimated very well when subpopulation "estimation groups" were formed properly. Using hydroelectric generation data by establishment entity from a previous annual census as the main regressor, missing y-value data were estimated for nonsampled members of the subpopulations in a monthly sample survey program. One region-of-interest was referred to as the Pacific Contiguous Census (Bureau) Region, which consists of these States: California, Oregon, and Washington. However, rainfall patterns are different for those States. An examination of a map produced by the US National Oceanic and Atmospheric Administration's National Climate Data Center (NOAA/NCDC) shows that it is better to group California with Nevada, and group Oregon and Washington with Idaho. This was done, and small area estimation by State began being performed, producing monthly estimates of hydroelectric generation totals by State, along with estimates of relative standard errors for these estimated (sub)totals. (Example 2 in Knaub(1999), pages 23 to 26, provides more details.)

However, at one point a few years later, for a few monthly publications, the estimated totals for hydroelectric generation for California were substantially degraded, which became somewhat obvious. It was then discovered that a software change had been made, inadvertently ignoring the argument above, and grouping the State data inappropriately. Once the software was returned to using the proper grouping, as specified in Knaub(1999), the problem was solved.[6] In this, and more complicated data requirement cases, model-based ratio estimation has performed very well for decades.

[6] In 1999, the US Energy Information Administration (EIA), because of Government budget constraints, but a large public appetite for more energy data, promised to increase publication of official statistics by way of a great many new categories of electric power generation and related fuel data, but without an appreciable increase in monthly sampling. To attempt to meet this goal, which management did not understand was not automatic, with many thousands of new sub-aggregate values to be published in categories for which there would often not be anything close to sufficient data, the author quickly developed the system touched upon in Knaub(1999), which was flexible to multiple problems in the rapid and frequent production of official statistics. The hydroelectric generation example was shown for simplicity in Knaub(1999), as other fuel-related cases are more complicated. Further, another multiple regression version of model-based ratio estimation was later developed to consider fuel switching by the population members, occurring between the annual census data collection, and monthly sample data collections. See discussion of this electric power plant fuel-switching problem in Knaub(2016a), on pages 23 and 24.

In addition, the reduction in variance over simple random sampling (SRS) is typically huge! See Knaub(2013). Sampling balanced such that $\sum_{i=1}^{n} x_i /n$ is approximately equal to $\sum_{i=1}^{N} x_i /N$ may reduce bias, as long as missed large members of the population are not important, but variance for balanced sampling will be about as bad as for SRS. Further, as noted below, one generally has more than one variable of interest on a survey, so the sample will generally not be optimal for all such variables/questions on a given survey. Thus, **typically when using a cutoff sampling approach, the sample will actually be a quasi-cutoff sample for each variable of interest, where not all of the largest members of the population for a given variable will be collected, and a scattering of smaller ones will be collected.** There is more below on multipurpose surveys, where more than one y-variable (question) is of interest.

Stratification by size, sigma, or category will often be helpful in either the model-assisted design-based case, or strictly model-based sampling and estimation ('prediction') case. Scatterplot graphics, as noted above, may be very helpful for this, as well as part of a data editing process.

For further information on quasi-cutoff sampling and estimation, see Knaub(2014a), and Knaub(2016b).

Regarding size measure for members of the population, Figure 2.1 on page 37 in Valliant, Dorfman, and Royall(2000) shows scatterplots which demonstrate the concepts of ignorable and nonignorable sample selection. There it shows the problem with using y-values instead of x-values as a measure of size. (For multiple regression, you can use predicted values of y, or any other combination of regressors, as a measure of size, but not the actual y-values.) That figure illustrates that selection is flawed if a cutoff sample is based on y-values, but not on x-values. Interestingly, a colleague, Joel Robert Douglas, was asked to explore a problem with a survey at the US Energy Information Administration on one occasion, and he traced the problem to this very issue.

4. Bias

Bias due to division of one of a pair of correlated randomly selected variables by another, estimated by a Taylor series, is explored explicitly for the case of simple random sampling in an excellent presentation in Sukhatme(1954), pages 138-146. In Cochran(1977), on pages 160 through 162, Section 6.8, "Bias of the Ratio Estimate," the treatment is somewhat different, and abbreviated, due, it appears, to the large range of topics covered well in Cochran(1977). On page 161, Cochran approximates bias for simple random sampling, as did Sukhatme, but with differences that might be described as follows:

(1) In Sukhatme(1954), it is shown how bias is eliminated when you have linear regression through the origin. If that is defined as homoscedastic, we know, for example, from Maddala(2001), page 208, that if the real relationship is heteroscedastic (as should be expected[7]), we still have unbiasedness. After page 143 in Sukhatme(1954), where he shows

[7] If we expect y to be zero when x is zero, we should often expect heteroscedasticity. That is, if the origin should be a point on a linear regression, would we expect numbers for y|x to include such possibilities for a prediction interval as 1,000,000 +/- 1,000, as well as 1,000 +/- 1,000, or even 5 +/- 1,000? Certainly as we approach the origin, prediction intervals should become

that unbiasedness means a linear relationship through the origin is required, he goes on to say, on page 144, that the estimated coefficient, b, which appears homoscedastic on page 143, can be replaced by $\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$, which in the model-based case occurs for $b = \left(\frac{\sum_{i=1}^{n} x_i y_i w_i}{\sum_{i=1}^{n} x_i^2 w_i}\right)$ when $\gamma = 0.5$, the classical ratio estimator, not when $\gamma = 0$, the case of homoscedasticity as used in 'ordinary' least squares regression. He shows, on page 144 that the expected value of $\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$ is $\frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$. Thus when any ratio estimate exactly applies, it is unbiased.[8]

(2) In Cochran(1977), on page 161, it can be seen that bias is reduced as $\bar{x}$ becomes closer to $\bar{X}$. This would be approximated by simple random sampling, especially with larger sample sizes, and in the model-based case indicates the usefulness of sampling balanced on the mean of the regressor. He also notes the need for a linear regression with a zero-intercept on page 158, and basically notes that the term "model-unbiased" applies when the expected sum of the estimated residuals is zero.

In Thompson(2012), near the bottom of page 102, he notes that the generalized (design-based) ratio estimator, which is written here as $\left(\frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}}{\sum_{i=1}^{n} \frac{x_i}{\pi_i}}\right)\left(\sum_{i=1}^{N} x_i\right)$, is also such that we have a ratio of two unbiased estimates, $\hat{R} = \left(\frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}}{\sum_{i=1}^{n} \frac{x_i}{\pi_i}}\right)$, which itself is not unbiased from a design-based perspective.

From a model-based perspective, we only need $b = \left(\frac{\sum_{i=1}^{n} x_i y_i w_i}{\sum_{i=1}^{n} x_i^2 w_i}\right)$, with $w_i = x_i^{-2\gamma}$. But the more general estimate for the 'slope' for a more general model-assisted design-based ratio

---

shorter, and skewed distortions to those intervals become apparent as we approach the origin, when we may not expect negative numbers, which may be nonsensical to an application.

[8] Three discrete values for γ are often used, though it is a continuous number, and an estimate of γ may be made, based on the data. (See, for example, Knaub(1997).) The three values for γ, as defined in this paper, which correspond to homoscedasticity, the CRE, and a mean of ratios, are, respectively, 0, 0.5, and 1.0. In Thompson(2012), on page 109, he describes those three cases, but refers to the homoscedastic case as regression through the origin. No matter the value of γ, these are all linear regressions through (really 'to') the origin. However, on page 127 in Valliant, Dorfman, and Royall(2000), they also appear to associate the term "through the origin" with a homoscedastic model. In Särndal, Swensson, and Wretman(1992), they use the term "ratio" when $\gamma = 0.5$ (actually 1, as γ is defined there), and all other cases are termed 'alternative ratios,' as shown in Section 7.3.4 of that book, where probability of selection weights are also used with regression weights. The latter weights are determined by the coefficient of heteroscedasticity, γ. - Another notable place where these three key values for γ are found is the second half of page 149 in Sukhatme(1954). For the Greek population study mentioned earlier, here we say that $\gamma = 1$ was used to estimate variance, though page 11 states that only probability-based sampling is considered in that book. For optimal design-based sampling, $\gamma = 1$ would require probability proportional to size (PPS, or PPx) sampling, which is often used. There is a reference to using the mean of individual sample member ratios for estimation with PPS/PPx sampling on page 167 in another old textbook, Yates(1949). – Also, please be aware that the notation in Särndal, Swensson, and Wretman(1992) is such that $\gamma = 1$ here is $\gamma = 2$ there.

estimator is $b = \left(\frac{\sum_{i=1}^{n} x_i y_i w_i/\pi_i}{\sum_{i=1}^{n} x_i^2 w_i/\pi_i}\right)$, where this is written in Section 7.3.4, "Alternative Ratio Models," in Särndal, Swensson, and Wretman(1992), with regression weights, w_i, defined as the inverse of those here.

The design-based bias, which Sukhatme(1954) shows to completely 'vanish' with a linear regression through the origin, will alternatively be reduced with larger sample sizes, n. Removal of design-based bias is also possible by modification of the ratio estimator, or sample selection method, as noted in Cochran(1977), pages 174-175, citing various sources. But the statement by Sukhatme(1954), on page 143, that when we have a linear regression with a zero intercept, under the design-based concept, bias disappears, is quite the bridge to a model-based concept of bias.

Bias in the model-based case is also due to assuming linear regression through the origin. Only y_i is considered to be a random variable under the strictly model-based approach. Of importance is the *conditional* distribution of y_i on each x_i. As noted above, a specific degree of heteroscedasticity is associated with the Classical Ratio Estimator (CRE), and other coefficients of heteroscedasticity may be more appropriate, either as estimated from the data, as in Knaub(1997), or perhaps reduced to guard against disproportionately large nonsampling error for smaller members of the population, as noted in Knaub(2009), pages 5 and 6, and Knaub(2010), page 4, but *b* remains unbiased (Maddala(2001), page 208).

For the model-based approach, we consider the model to be correct. However, bias due to 'model-failure' can be examined *post hoc*. The $y_i - bx_i$ values collected for the sample are now estimated residuals for a model, and we are more generally considering $(y_i - bx_i)w_i^{0.5}$, rather than considering the collected data under a randomized probability of selection design. Note that although $b = \hat{R}$, we generally see $y_i - \hat{R}x_i$ written for (model-assisted) design-based sampling and estimation, and $y_i - bx_i$ is used for (strictly) model-based sampling and prediction.[9] We consider "model-unbiasedness" in the latter case, as mentioned in Cochran(1977), which means we look for models where the sum of the estimated residuals has an expected value of zero. For the CRE that sum is always exactly zero. Bias occurs because of "model-failure," because the model is not exactly correct. (See Knaub(2010).) Test data and graphical analyses may be used to study bias, *post hoc*, as in Knaub(2001), and some of Knaub and Douglas(2010).[10] For official statistics which are derived from finite population samples at regular intervals, various methods may be used, including subject matter knowledge, to validate results.

[9] In the former case, the x_i are "auxiliary" data, and in the latter case, they are "regressor" data.
[10] Barely mentioned in Knaub and Douglas(2010) was an extensive, painstaking study, by Brett Foster and Lisa Guo, summer interns at the US Energy Information Administration (EIA), in the Joint Program in Survey Methodology (JPSM) Junior Fellow interns program. They compared cases of estimated annual totals found from 12 monthly samples, whose estimated totals were published in close to 'real time,' to the later collected annual census derived totals, to compare these differences with estimated relative standard errors.

5. Variance

Under the design-based (probability of selection) approach, there is a randomized selection of the (x_i, y_i), to thus obtain a randomized selection of values for $\left(y_i - \hat{R}x_i\right)^2$ to be used for estimation of sigma.

Under the model-based (regression/prediction) approach, we select the y_i for the sample, as Cochran(1977) notes at the bottom of page 158, only with regard to the x_i values, $y_i|x_i$, with sigma then based on $(y_i - bx_i)^2 w_i$. For the classical ratio estimator (CRE), this is $(y_i - bx_i)^2 x_i^{-1}$. This becomes part of a development on page 159, Cochran(1977), using a Lagrange multiplier, where interestingly, Cochran finds that minimizing variance for the CRE has him using $\left(y_i - \hat{R}x_i\right)^2 x_i^{-1}$. (See Cochran (1977), page 159, equation 6.27.) For more on this, see Knaub(2016a), pages 15 through 19, which compares design-based variance for simple random sampling with regression model-based variance, with heavy reference to Cochran(1977).

In model-based ratio estimation/prediction, the y_i, and the total to be estimated, are random variables from a superpopulation. (See Cochran(1977), page 158.) There is a component to the variance due to the "irreducible error" (Fortmann-Roe(2012))[11], which is the part for sigma only, not the regression coefficient(s), which is (are) also based on sigma. Thus we are looking at an analysis of variance in that there are at least two components to the variance of the prediction error. This relates to the second term in Cochran's variance estimate/prediction for the predicted total of equation 6.26, on page 159, which appears to give rise to the "nonsample part" in Valliant, Dorfman, and Royall(2000), pages 130-134. Using compatible notation here, what Valliant, Dorfman, and Royall call $V\left(\hat{Y}_R - Y_R\right) = V\left(\hat{Y}_R\right) + V(Y_R)$, Cochran just calls $V\left(\hat{Y}_R\right)$.

Note that Section 5.1.2 "Variance Estimators for the Ratio Estimator," in Valliant, Dorfman, and Royall(2000), they show what happens when you introduce the 'hat matrix' to account for individual data point influences, and do not use regression weights. However, in the context of practical work experience for production of official energy statistics for many small populations with small samples, this author has found it very robust to use weighted least squares with an underestimated coefficient of heteroscedasticity. When the coefficient of heteroscedasticity, γ, for an establishment survey was estimated for energy data at the US Energy Information Administration (EIA), it was very often the case that $0.8 < \gamma < 0.9$, and it was virtually always in the range $0.5 < \gamma < 1.0$, as expected in the explanation by Brewer(2002) on page 111. Experience at the EIA showed generally good results using $\gamma = 0.5$, thus using the classical ratio estimator (CRE), which appeared robust to nonsampling error issues for smaller respondents. (Sometimes further steps were necessary when the data are very 'dirty.' See Knaub(2009), top of page 6, but generally the CRE performed well.)

With regard to analysis of variance, there is an interesting relationship between the usual analysis-of-variance method, and the partitioning of the estimated variance of prediction error. On page 293 of Walpole and Myers(1972), they use SST = SSR + SSE to represent

[11] See also Stansbury(2013) for another illustration of the bias-variance concept involved.

the partitioning of the total sum of squares into a part based on the model and a part based on residuals only, *i.e*, a part that is "explained" by the model and a part that is not, where this latter part is a sum of squared "errors." This has the "feel" of a probability of selection (design-based) approach, because it is based on the unconditional distribution of y-values, even though users may often never consider the appropriateness of the sampling methodology, or even the sample size.[12] At any rate, let us compare SST = SSR + SSE to the analogous "analysis of variance" of sorts we may see from examining the estimated variance of the prediction error of totals, shown, for example, on page 19 of Knaub(2009):

$V_L^*(T^* - T) = \sigma_{e_0}^{*2} \sum_{i=n+1}^{N} x_i^{2\gamma} + V^*(b)\left(\sum_{i=n+1}^{N} x_i\right)^2$, where $V^*(b) = \sigma_{e_0}^{*2} / \sum_{i=1}^{n} x_i^{2-2\gamma}$, and $\sigma_{e_0}^{*2} = \sum_{i=1}^{n} e_{0i}^2/(n-1)$, for the ratio estimator based on $y_i = bx_i + e_{0i}x_i^{\gamma}$, with $b = \frac{\sum_{i=1}^{n} x_i^{1-2\gamma} y_i}{\sum_{i=1}^{n} x_i^{2-2\gamma}}$.

There, that is referred to as "regression through the origin," though it is for any value of the coefficient of heteroscedasticity, including zero, where Thompson(2012), page 109 labeled that as the "regression-through-the-origin estimator."

As with SST = SSR + SSE, when we write $V_L^*(T^* - T) = V^*(b)\left(\sum_{i=n+1}^{N} x_i\right)^2 + \sigma_{e_0}^{*2} \sum_{i=n+1}^{N} x_i^{2\gamma}$, we are partitioning into a part based on the model, and a part based on random error alone, respectively. However, here we are considering the conditional distribution of y given x, as shown on page 18 in Maddala(2001) as $f(y|x)$. For $y|x$, we do not look at $\bar{y}$, as that only matters to the unconditional distribution of y. But we still partition variance into a part for the model coefficient parameter(s), and one based on the "irreducible error"[13] noted above. So another way to look at model 'fit' for ratio predictions would be to consider $V^*(b)\left(\sum_{i=n+1}^{N} x_i\right)^2 / V_L^*(T^* - T)$, instead of SSR/SST = r^2. However, we should also note that the unconditional approach, SST = SSR + SSE, considers the sample only, and may even be used for an infinite population, whereas the conditional ('fully' model-based) approach for totals considered here, though we could have looked at $V_L^*(y_i^* - y_i) = x_i^2 V^*(b) + \sigma_{e_0}^{*2} x_i^{2\gamma}$, is for a finite population.[14] As with any variance estimation for a population, $V_L^*(T^* - T)$ makes use of an estimate of sigma from the n members of the sample, and then applies it to the N-n members of the population which are not in the sample (analogous to applying a finite population correction factor). SSE, really $\sum_{i=1}^{n}(y_i - bx_i) = \sum_{i=1}^{n} e_i$, compares to $\sigma_{e_0}^{*2} x_i^{2\gamma}$. For $V_L^*(T^* - T)$, we are looking at the N-n population members not in the sample.

[12] This may often mislead users of statistical software when they see, for example r = 0.98 on a graph with two data points, and think all is fine.

[13] Being "irreducible" is contingent upon having the completely "correct" model, which is not possible in practice.

[14] Note that since we have $y_i|x_i$, when we consider heteroscedasticity, as we do here, if we used $x_i^2 V^*(b)/V_L^*(y_i^* - y_i)$ in place of r^2, it would be different for each member of the population. Also note that whether we use $V^*(b)(\sum_{i=n+1}^{N} x_i)^2 / V_L^*(T^* - T)$ or $x_i^2 V^*(b)/V_L^*(y_i^* - y_i)$, $\sigma_{e_0}^{*2}$ 'cancels' from the numerator and denominator, so we are left with functions of x only. Thus, like r, we would have a measure of dubious stand-alone worth.

6. Multiple Regression

In Cochran(1977), on page 184 through 186, section 6.20, "Multivariate Ratio Estimates," he presents the case of design-based ratio estimation with multiple auxiliary variables, referencing an article in *Biometrika*, 1958, by Ingram Olkin. Today, "multivariate" might mean multiple dependent variables, rather than multiple auxiliary (or regressor/independent) variables, which is what is meant here. On page 186, Cochran notes that use of a second (or more) auxiliary variable(s) can substantially improve "precision," where he is comparing variance estimates. In statistical learning, the bias-variance tradeoff tells us that increased model complexity, such as more regressor variables, generally leads to less bias, but more variance. (Note the impact, however, of *estimated* sigma on *estimated* variance of the prediction error, such as discussed in Knaub(2016a), on pages 23 and 24.)[15] Here, the design-based multiple auxiliary variable ratio estimator, as described on page 185 in Cochran(1977), is actually a linear combination of the usual single variable ratio estimators, not a multiple regression in the usual sense. This is the (at least now) familiar technique used in different types of applications, where weights are assigned to more than one estimation (prediction), such that the weights will sum to unity. These weights, based on a covariance matrix, in the case of two auxiliary variables, are described primarily on page 185 in Cochran(1977).

In the model-based case of actually using multiple regression, we may say, for example, for two regressors, that $y_i = b_1 x_{1_i} + b_2 x_{2_i} + e_{0_i} w_i^{-0.5}$ where $w_i = z_i^{-2\gamma}$. Here, z_i is a measure of size, analogous to cluster size, as x_i is possible to be considered to be a cluster size for one regressor. $\gamma = 0.5$ corresponds to the one-regressor model-based Classical Ratio Estimator (CRE). For $\gamma = 0.5$, the $e_{0_i} w_i^{-0.5} = e_{0_i} z_i^{0.5} = e_i$ always sum exactly to zero, as with the usual model-based CRE. (See Särndal, Swensson, Wretman(1992), Example 6.5.1, page 232,[16] with regard to multiple linear regression.)

7. Multipurpose Surveys

Data collection is generally done for more than one y variable of interest. A design-based approach means one set of π_i (probability of selection) values provides the optimal sampling, and best estimation, for one y-variable of interest. That may often be far from ideal for the estimation phase for numerous other y-variables. Cassel, *et.al.*(1977/1993), pages 107, and 150 notes this problem, and suggests on page 150 that a compromise size measure needs to be used. Holmberg(2007) considers a compromise based on modeling this problem with regard to unequal probability sampling. Page 107 in Cassel, *et.al.*(1977/1993) notes work by J. N. K. Rao, published in 1966, which they say considered

[15] Aside: Bias-variance tradeoff considerations in Statistical Learning tell us that generally when we increase the complexity of a model, such as the number of regressors, we decrease bias and increase variance. But because *estimated* sigma is part of the *estimated* variance of the prediction error, overestimation of sigma may 'absorb' the assumption of linearity through the origin, to a degree, and thus the estimated variance of the prediction error can be a good overall measure of accuracy, though it somewhat conflates bias and variance. That is, by adding a needed regressor, *estimated* variance of the prediction error may be reduced due to the way sigma is estimated. (See discussion of electric power plant fuel switching problem in Knaub(2016a), on pages 23 and 24.)

[16] Thank you to Phil Kott for noting this page of that reference, on a different occasion, in another context.

when y and x may be unrelated. In Thompson(2012), page 104, he notes that for cases where y is not linearly related to either the probabilities of selection, nor to an "auxiliary variable," we may use a generalized ratio estimator, as $\left(\frac{\sum_{i=1}^{n}\frac{y_i}{\pi_i}}{\sum_{i=1}^{n}\frac{x_i}{\pi_i}}\right)\left(\sum_{i=1}^{N}\boldsymbol{x_i}\right)$, with the $\boldsymbol{x_i}$ all set equal to 1, thus using $\left(\frac{\sum_{i=1}^{n}\frac{y_i}{\pi_i}}{\sum_{i=1}^{n}\frac{1}{\pi_i}}\right)N$. Thompson(2012), pages 103 and 104, explains, though not stated this way, that this would have saved the circus statistician's job in Basu's (in)famous elephant fable.[17] However, it would seem we could often do better with a strictly model-based approach.

Often, in survey statistics, we have one or more auxiliary variables/regressors available for each y-variable for which we wish to use ratio estimation. (This occurs frequently in the area of Official Statistics.) If we stratify or post-stratify the population for one reasonable model per subpopulation, then a strictly model-based prediction approach may be used, and applied for estimation (really prediction), one y-variable at a time, often to great advantage. That is, for the model-based approach, the estimation phase is customized by y-variable and stratum. A quasi-cutoff compromise sample may be used. The customized estimate for each y variable does not depend upon one size measure. In the design-based approach, when confined to one size measure, the set of π_i values, for all y variables, this could provide very inaccurate results in the estimation phase, without some kind of adjustment.

Calibration for multivariate y across the same vector of auxiliary variables (which is a model-assisted, probability design-based method) can be another, more complex solution for dealing with multipurpose surveys, under favorable circumstances. See Chambers and Clark(2012), page 141. However, even when appropriate for a practical situation, Chambers and Clark(2012) page 143 notes that this can be inefficient. In an application to Official Statistics, where many small populations are sampled, and the same survey will have many questions relating to basically different populations, for which a common set of useful auxiliary variables is not the case, this would not appear to be helpful.

8. Closing Remarks

The classical ratio estimator (CRE) is *cosmetically calibrated*, so under both design-based and model-based concepts, the estimators are identical.

For the design-based ratio estimator, we sample y_i, but effectively sample $e_i = y_i - \hat{R}x_i$, where $\hat{R} = \frac{\bar{x}}{\bar{y}}$, and we employ a version of an expansion factor, because we assume we have a "representative" set of e_i used in variance estimation. How the $\bar{x}$ and $\bar{y}$ are estimated depends upon the design, and we are considering the unconditional distributions of *x*, and *y*. Simple random sampling (SRS) is suboptimal due to heteroscedasticity for e_i. One may stratify by size of e_i, or use unequal probability sampling. Note that stratification by size is essentially a more granular version of unequal probability sampling.

For prediction from model-based ratio estimators, there is either a predicted/modeled or an observed value for each member of the population, for which we obtain a sum. (Note: The

[17] Note that Brewer(2002) is subtitled "Weighing Basu's Elephants."

sum of all estimated residuals for the model-based CRE is always exactly zero, and for the alternative ratio estimator where $b = \left(\sum_{i=1}^{n} \frac{y_i}{x_i}\right)/n$, the estimated random factors of the estimated residuals always sum to zero, as shown in Knaub(2005).) Further, the sum of the estimated variances for each prediction error differs from the estimated variance of the prediction error for the total, due to the difference between a sum of squares and the square of a sum, with regard to b, which is not applicable to the σ, "irreducible" part. (See Knaub(1999), pages 4 and 5. Erratum: On page 4 there, "multivariate" should be replaced by "multiple linear regression.")

Bias from the perspective of simple random sampling (SRS), as shown in Sukhatme(1954), is because y_i and x_i are correlated random variables. Because both y_i and x_i are random variables based on probability of selection, the derived estimator is only unbiased when the relationship is linear through the origin, as shown in Sukhatme(1954), pp 138-144. Thompson(2012), page 102, also notes the design unbiasedness of the numerator and denominator of what amounts to $\hat{R}$ in the generalized (design-based) ratio estimator, but a biased result for $\hat{R}$ is also true, thus generalizing to other probability-based designs.

Bias (model-failure) for model-based prediction is also due to failure of a linear relationship between y_i and x_i, which must include the origin. Only y_i is considered a random variable in the model, not x_i. y_i is conditional on x_i, and thus its conditional distribution (that of e_i) is very different from the (unconditional) y_i population distribution.

Note on page 102 of Thompson(2012), that while discussing the design-based approach, he notes that ratio estimation is advantageous when the variance of the $y_i - Rx_i$ is ("much") smaller than the variance of the y_i.

Ratio estimation has many uses. In Lohr(2010), on page 180, she provides an example of the use of a ratio when a mean per cluster is less variable than a total per cluster, due to unequal cluster size sampling. A common theme regarding ratio estimates holds here: We may trade addition of a small bias for a large decrease in variance.

As noted in Knaub(2017), whether intended or not, use of a ratio estimator with γ>0, say the CRE (thus γ=0.5), using either a probability of selection or model-based approach, we are giving more emphasis (weight) to the smaller (x-value) members of a population.

9. Conclusions

This paper addresses various forms of ratio estimators often used for establishment and other sample surveys. Concentration has been on application to Official Statistics. The two main philosophies considered involve sampling and estimation based on (1) probability of selection, or (2) regression modeling/'prediction' – or a combination of the two. Biases and variances for these two philosophies, though the philosophies are different in concept, have much in common. It is important to understand those measures of accuracy when considering results.

The primary difference in concept is that of sampling $y_i - \hat{R}x_i$ under a design-based probability-of-selection scheme, where we consider the unconditional distribution of the y_i, or do we collect the same statistic, but write it as $y_i - bx_i$, and consider the conditional

distribution of $y_i|x_i$. In both cases unbiasedness is achieved when linear regression with a zero intercept is the case. Thus for all practical purposes, we can think of all ratio estimation as a model-driven procedure. The use of a design employing probability of selection is more of a safeguard against model-failure, where variance estimates are based on $\left(y_i - \hat{R}x_i\right)^2$ rather than $(y_i - bx_i)^2 w_i$, as an attempt at robustness. However, the very nature of survey sampling with a correlated auxiliary/regressor variable, linearly related to each other with a zero-intercept, which is very often the case, tells us that there will be a strong degree of heteroscedasticity, for which simple random sampling is highly inefficient, and simple balanced sampling produces that same problem as well.

As always, stratifications can be by category or size, and can be very helpful under either philosophy.

Putting prediction intervals around each line can help us determine which groups to separate, and what may be modeled together. See Figure 3, page 13 in Knaub(2012), for example. (Note that near the origin, if negative numbers are not valid, those intervals would be very noticeably non-"normal," contrary to the illustration.)

Appendix: Overview

The following overview was a 'handout' which accompanied a poster presented in Baltimore, Maryland, USA, on August 1, 2017, at the Joint Statistical Meetings. That poster is found at the following location, and was presented under the auspices of the American Statistical Association's Survey Research Methods Section:

https://www.academia.edu/33469291/Poster_for_Comparison_of_Model-Based_to_Design-Based_Ratio_Estimators_

Also, see
https://www.researchgate.net/publication/317523999_Poster_for_Comparison_of_Model-Based_to_Design-Based_Ratio_Estimators

Survey inference depends on whether you rely upon a reasonable model, or upon randomized sample selection, or a combination of both. For ratio estimators, a simple model is either considered directly in the model-based case, or in an indirect sense in a design-based process. – See reference list handout, which includes Thompson(2012). Although the model-based and design-based approaches to ratio estimators are quite different in philosophy, they have much in common:

bias:
both assume a linear relationship "*through*" the origin, *or else* they are biased:
for the model-based case, bias is considered "model-failure," but
for the design-based case, it is due to the use of "$\hat{R}$," a ratio constructed from random variables y and x, random in the sense of selection, and compared to the standard error, this bias becomes smaller with larger sample sizes (particularly good resource: Sukhatme(1954))

variance:
both are based on y-bx (with $b = \hat{R}$), not just based on y,
for the model-based case, y-bx is an estimated residual, and the estimated variance of the prediction error may be analyzed,
for the design-based case, the variance of $\hat{R}$ is constructed from the estimated standard errors of y and x, and their estimated covariance –
note: Variance for skewed populations can be very high for design-based sampling without using optimal unequal probability sampling, which will not be optimal for all y-variables of interest*

Note: The "generalized ratio estimator" incorporates Horvitz-Thompson estimators, and "Alternative Ratio Models," Särndal, Swensson, and Wretman(1992) , pp. 254, 255, 246, further incorporate regression weights. –
[The smallest x-selected units may still be problematic from a data collection quality perspective, but the strictly model-based approach only uses the regression weights. Very small problematic data may better be predicted.]

The impact of heteroscedasticity is felt in each case:
in the model-based case, this means using weighted regression, and
in the design-based case, it means that simple random sampling is not optimal: need unequal probability sampling

note: This is related to unit size in cluster sampling (Cochran(1977), p 243).

Related to cluster sampling, with cluster size x_i (or perhaps predicted-y_i for multiple regression), and within-cluster variance proportional to $x_i^{2\gamma}$. See Cochran(1977), p. 243.

Result: $\sigma_i^2 \propto x_i^{2\gamma}$: $0.5 \leq \gamma \leq 1$, Brewer(2002), p 111

Stratification can aid either method immensely:
modeling should be applied by portion of the population, for which each portion is approximately governed by a given model (a consideration which also applies to 'borrowing strength' for small area estimation), and
similarly, for probability of selection, design-based sampling, reduced within stratum variance and increased mean differences between strata is helpful

Note: In the design-based case, a "representative" data selection may be accomplished in an overall sense. For the model-based case, without properly stratifying the population, and/or a "balanced" sample, this may be a problem. However, for a highly skewed population, a sample balanced on the mean for x, as with a simple random sample, will have a *huge* efficiency disadvantage.

*For multipurpose surveys – and almost every survey does collect data on more than one item – a model-based ratio estimator will provide better weighting … no Basu's Elephant problem!
– The design-based estimator can be adjusted, but the model-based case is straightforward.

Acknowledgements

Thank you to Ken Brewer, friend and mentor, for past long-term help and encouragement.

Also, thank you to those researchers on ResearchGate who answered my requests for suggestions for more textbooks to consult.

This paper and associated documents may be found by going through

https://www.researchgate.net/profile/James_Knaub/contributions
and/or
https://independent.academia.edu/JamesKnaub

jamesRknaub@gmail.com

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. (2002), *Combined survey sampling inference: Weighing Basu's elephants*, Arnold: London and Oxford University Press.

Brewer, K.R.W. (2014), "Three controversies in the history of survey sampling," *Survey Methodology*, (December 2013/January 2014), Vol 39, No 2, pp. 249-262. Statistics Canada, Catalogue No. 12-001-X. http://www.statcan.gc.ca/pub/12-001-x/2013002/article/11883-eng.htm

Cassel, C.M., Särndal, C.E., and Wretman, J.H.(1977, 1993), *Foundations of Inference in Survey Sampling*, Wiley 1977, Krieger Publishing Company 1993.

Chambers, R.L., and Clark, R.G.(2012), *An Introduction to Model-Based Survey Sampling with Applications*, Oxford University Press.

Cochran, W.G.(1977), *Sampling Techniques*, Third Edition, John Wiley & Sons.

Deming, W.E. (1938, 1943, 1964), *Statistical Adjustment of Data*, 1964 corrected Dover republication of the 1943 John Wiley & Sons, Inc. publication.

Deming, W.E. (1950, 1966), *Some Theory of Sampling*, John Wiley & Sons, Inc., Republished by Dover Publications, unaltered and unabridged.

Douglas, J.R.(2013), "Efficiently Utilizing Available Regressor Data Through a Multi-Tiered Survey Estimation Strategy," *InterStat*, September 2013, http://interstat.statjournals.net/YEAR/2013/abstracts/1309001.php?Name=309001

Douglas, J.R., and Knaub, J.R., Jr.(2010), "Using Predicted Explanatory Variables and Their Effects on Variance Estimations under Weighted Least Squares Regression," *InterStat*, February 2010, http://interstat.statjournals.net/, found at https://www.academia.edu/16492126/Using_Predicted_Explanatory_Variables_and_Their_Effects_on_Variance_Estimations_under_Weighted_Least_Squares_Regression

Fairfield Smith, H.(1938). An empirical law describing heterogeneity in the yields of agricultural crops. *The J. Agri. Sci.*, 28, 1-23.

Fortmann-Roe, S.(2012), "Understanding the Bias-Variance Tradeoff," http://scott.fortmann-roe.com/docs/BiasVariance.html

Hansen, M.H., Hurwitz W.N. and Madow, W.G. (1953), *Sample Survey Methods and Theory*, Volume II, Theory, Republished 1993, Wiley, New York.

Holmberg, A.(2007), "Using Unequal Probability Sampling in Business Surveys to Limit Anticipated Variances of Regression Estimators," *Proceedings of the Third International Conference on Establishment Surveys* (Montreal, Quebec, Canada), American Statistical Association, pp. 550-556.
http://www.amstat.org/meetings/ices/2007/proceedings/ICES2007-000162.PDF

Kish, L.(1965, reprinted 1995), *Survey Sampling*, Wiley.

Knaub, J.R., Jr. (1997), "Weighting in Regression for Use in Survey Methodology," *InterStat*, April 1997, http://interstat.statjournals.net/. (Note shorter, but improved version in the 1997 Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 153-157.)
https://www.researchgate.net/publication/263032446_Weighting_in_Regression_for_Use_in_Survey_Methodology

Knaub, J.R., Jr.(1999), "Using Prediction-Oriented Software for Survey Estimation," *InterStat*, August 1999, http://interstat.statjournals.net/, partially covered in "Using Prediction-Oriented Software for Model-Based and Small Area Estimation," in *Proceedings of the ASA Survey Research Methods Section*, 1999, and partially covered in "Using Prediction-Oriented Software for Estimation in the Presence of Nonresponse," presented at the *International Conference on Survey Nonresponse*, 1999.
https://www.researchgate.net/publication/261586154_Using_Prediction-Oriented_Software_for_Survey_Estimation

Knaub, J.R., Jr. (2001), "Using Prediction-Oriented Software for Survey Estimation - Part III: Full-Scale Study of Variance and Bias," *InterStat*, June 2001, http://interstat.statjournals.net/. (Note another version in ASA Survey Research Methods Section proceedings, 2001.)
https://www.researchgate.net/publication/261588075_Using_Prediction-Oriented_Software_for_Survey_Estimation_-_Part_III_Full-Scale_Study_of_Variance_and_Bias

Knaub, J.R., Jr. (2005), "The Classical Ratio Estimator (Model-Based)," *InterStat,* October 2005, http://interstat.statjournals.net/
https://www.researchgate.net/publication/261474011_The_Classical_Ratio_Estimator_Model-Based

Knaub, J.R., Jr.(2009), "Properties of Weighted Least Squares Regression for Cutoff Sampling in Establishment Surveys," *InterStat,* December 2009, http://interstat.statjournals.net/.
https://www.researchgate.net/publication/263036348_Properties_of_Weighted_Least_Squares_Regression_for_Cutoff_Sampling_in_Establishment_Surveys

Knaub, J.R., Jr.(2010), "On Model-Failure When Estimating from Cutoff Samples," *InterStat,* July 2010, http://interstat.statjournals.net/.
https://www.researchgate.net/publication/261474154_On_Model-Failure_When_Estimating_from_Cutoff_Samples

Knaub, J.R., Jr.(2011), "Ken Brewer and the Coefficient of Heteroscedasticity as Used in Sample Survey Inference," *Pakistan Journal of Statistics*, Vol 27, No 4, 397-406. https://www.researchgate.net/publication/261596397_Ken_Brewer_and_the_coefficient_of_heteroscedasticity_as_used_in_sample_survey_inference

Knaub, J.R., Jr.(2012), "Use of Ratios for Estimation of Official Statistics at a Statistical Agency," *InterStat,* May 2012, http://interstat.statjournals.net/. https://www.researchgate.net/publication/261508465_Use_of_Ratios_for_Estimation_of_Official_Statistics_at_a_Statistical_Agency

Knaub, J.R., Jr.(2013), "Projected Variance for the Model-Based Classical Ratio Estimator: Estimating Sample Size Requirements," *Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 2885-2889. https://www.researchgate.net/publication/261947825_Projected_Variance_for_the_Model-based_Classical_Ratio_Estimator_Estimating_Sample_Size_Requirements

Knaub, J.R., Jr.(2014a), "Efficacy of Quasi-Cutoff Sampling and Model-Based Estimation For Establishment Surveys and Related Considerations," *InterStat,* January 2014, http://interstat.statjournals.net/. https://www.researchgate.net/publication/261472614_Efficacy_of_Quasi-Cutoff_Sampling_and_Model-Based_Estimation_For_Establishment_Surveys_and_Related_Considerations

Knaub, J.R., Jr.(2014b), "Quasi-Cutoff Sampling and Simple Small Area Estimation with Nonresponse," *InterStat,* May 2014, http://interstat.statjournals.net/. https://www.researchgate.net/publication/262066356_Quasi-Cutoff_Sampling_and_Simple_Small_Area_Estimation_with_Nonresponse

Knaub, J.R., Jr.(2015), "Note on Ratio and Chain Ratio-Type Estimators: Gamma and Alpha Coefficients," unpublished Technical Note, only on ResearchGate. https://www.researchgate.net/publication/272485945_Note_on_Ratio_and_Chain_Ratio-Type_Estimators_Gamma_and_Alpha_Coefficients

Knaub, J.R., Jr.(2016a), "Prediction for Finite Populations: Cutoff Sampling and Related Issues," *ResearchGate*, https://www.researchgate.net/publication/301285520_Prediction_for_Finite_Populations_Cutoff_Sampling_and_Related_Issues

Knaub, J.R., Jr.(2016b), "When and How to Use Cutoff Sampling with Prediction," unpublished Method, only on ResearchGate. https://www.researchgate.net/publication/303496276_When_and_How_to_Use_Cutoff_Sampling_with_Prediction

Knaub, J.R., Jr.(2017), "Quasi-Cutoff Sampling and the Classical Ratio Estimator: Application to Establishment Surveys for Official Statistics at the US Energy Information Administration," Math/Stats Lunch presentation, to be delivered September 27, 2017 and made available under https://www.researchgate.net/profile/James_Knaub/contributions

Knaub, J.R., Jr., and Douglas, J.R.(2010), "Cutoff Sampling and Estimation for Establishment Surveys," Slides for a seminar presented at the U.S. Energy Information Administration, June 2010, found at
https://www.academia.edu/16434331/Cutoff_Sampling_and_Estimation_for_Establishment_Surveys
and at
https://www.researchgate.net/publication/263927238_Cutoff_Sampling_and_Estimation_for_Establishment_Surveys,
DOI: 10.13140/RG.2.1.1001.3282

Lohr S.L.(2010), *Sampling: Design and Analysis*, Second Edition, Brooks/Cole.

Maddala, G.S.(2001), *Introduction to Econometrics*, Third Edition, John Wiley & Sons.

Murthy, M.N.(1967), *Sampling Theory and Methods*, Statistical Publishing Society.

Raj, D.(1968), *Sampling Theory*, McGraw-Hill.

Rossel, R.A.V., Brus, D.J., Lobsey, C., Shi, Z., McLachlan G.(2016), "Baseline estimates of soil organic carbon by proximal sensing: Comparing design-based, model-assisted and model-based inference," Geoderma, 265, Open Access, pp. 152-163, Science Direct, Elsevier. http://www.sciencedirect.com/science/article/pii/S0016706115301312.

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

Stansbury, D.(2013), "Model Selection: Underfitting, Overfitting, and the Bias-Variance Tradeoff,"
https://theclevermachine.wordpress.com/2013/04/21/model-selection-underfitting-overfitting-and-the-bias-variance-tradeoff/, *Topics in Computational Neuroscience & Machine Learning*.

Sukhatme, P.V. (1954), *Sampling Theory of Surveys with Applications*, The Iowa State College Press.

Thompson, S.K.(2012), Sampling, Third Edition, John Wiley & Sons.

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

Walpole, R.E., and Myers, R.H.(1972), *Probability and Statistics for Engineers and Scientists*, The Macmillan Company.

Warren, B.(2004), "Design-Based, Model-Based, and Model-Assisted Inference: A Tutorial," A discussion paper produced for the Vegetation Resources Inventory Section, Resource Information Branch, Ministry of Sustainable Resource Management, British Columbia, Canada, 30 July 2004, previously available at
https://www.for.gov.bc.ca/hts/vri/technical/technical/design_model_tutorial_final.pdf

Yates, F.(1949), Sampling Methods for Censuses and Surveys, Charles Griffin & Company Limited, London.

Note: Other resources of interest may be found in a bibliography at the following location: https://www.researchgate.net/publication/317914104_Handout_Bibliography_for_Comparison_of_Model-Based_to_Design-Based_Ratio_Estimators_Poster.
DOI: 10.13140/RG.2.2.32164.07049