The Classical Ratio Estimator

Abstract:

The classical ratio estimator (CRE) is very simple, has a long history, and has a stunningly broad range of application, especially with regard to survey statistics, particularly establishment survey statistics. The CRE has a number of desirable properties, one of which is that the sum of its estimated residuals is always zero. It is easily extended to multiple regression, and a property shown in Sarndal, Swensson and Wretman (1992) may be used to indicate the desirability of this zero sum of estimated residuals feature when constructing regression weights for multiple regression. In the single regressor form, the zero sum of estimated residuals property is related to an interesting phenomenon expressed in Fox (1997). Finally, relationships of the CRE to some other statistics are also considered.

Introduction:

The classical ratio estimator is one case that results from the linear regression model, $y_i = \hat{\beta} x_i + e_i$, with a zero intercept. From the standard derivation of weighted least squares regression, when the intercept is fixed at zero, with one regressor, x_i, we have $\hat{\beta} = \sum_{i=1}^{n} w_i x_i y_i / \sum_{i=1}^{n} w_i x_i^2$. See, for example, Abdi (2003). The "error" being minimized in that reference is actually the sum of squared estimated residuals, which in the zero intercept case for weighted least squares linear regression is $\sum_i w_i \left(y_i - \hat{\beta} x_i\right)^2$. Taking the partial derivative with respect to $\hat{\beta}$, and setting it equal to zero, we find $2\hat{\beta}\sum_{i=1}^{n} w_i x_i^2 - 2\sum_{i=1}^{n} w_i x_i y_i = 0$, so that $\hat{\beta} = \sum_{i=1}^{n} w_i x_i y_i / \sum_{i=1}^{n} w_i x_i^2$. When $y_i = \hat{\beta} x_i + x_i^{\gamma} e_{0i}$, where e_{0i} is the random factor of the residual e_i, then the regression weight is $w_i = x_i^{-2\gamma}$. Further, if $\gamma = 0.5$, then $\hat{\beta} = \sum_{i=1}^{n} x_i^{-1} x_i y_i / \sum_{i=1}^{n} x_i^{-1} x_i^2$ $= \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i$. So, $\hat{\beta} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i$ results when $e_i = z_i^{\gamma} e_{0i}$, where $z = x$ in the case of one regressor, and γ, the coefficient of heteroscedasticity (Brewer (2002), page 87 and 111) is 0.5. Thus the CRE of the total is $\left[\sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i\right] \sum_{i=1}^{N} x_i$.

History and Ubiquity:

The CRE has historical precedence dating at least to 1820 when Pierre-Simon Laplace estimated the population of France "… by means of a ratio estimator (ratio of population to births during the preceding year)," Cochran (1978), page 3. Applications are diverse. Carlson, Coggins and Swanton (1998) discuss this work as applied to a modern salmon marking and recapture experiment. The least squares method in general can be traced back to Adrien-Marie Legendre and Johann Carl Friedrich Gauss, both publishing very early in the 1800s. The CRE is a special, but widely applicable case of least squares regression. From Cochran (1978), page 5, an even earlier application of an "… estimated ratio of population to births …" than that of Laplace was documented by John Graunt in 1662. Cochran (1978), page 7, went on to say, that Laplace made use of superpopulations, and that "So far as I know, this use of an infinite superpopulation in studying the properties of sampling methods was not reintroduced into sample survey theory until 1963, when Brewer (1963), followed by Royall (1970), applied it to the ratio estimator with the following model in the superpopulation

$$y = \beta x + \varepsilon \qquad \varepsilon \sim (0, \lambda x),$$

this being the model under which the ratio estimator is expected to perform well." This is the classical ratio estimator, CRE.

The ratio estimator is used in a number of settings. Its general applicability is evident in its widespread use. Improvements may be made under certain circumstances, such as in the concept of balanced sampling (Royall and Cumberland (1981)). Balanced sampling can lower bias, but Knaub (2002), pages 2-3, and 17-19, found that in highly skewed establishment surveys, extraordinary nonsampling error in the smallest responses, when small respondents are required to respond on too frequent a basis, can make it impractical to use such data, thus forcing the use of a cut-off sample. However, whether a balanced, model-based sample, a cut-off model-based sample, or randomization for a design-based or model-based sample is implemented, ratio estimation has often proved useful, and the classical ratio estimator has long-standing and continued appeal. Note such an unusual application as found in "Efficient Sampling Design in Audit Data," (Liu, Y., Batcher, M. and Scheuren, F. (2005)). There a special "setting" is described "… with the estimation method employing a ratio type method." Small area estimation is another application, likely a much more common one, although the ratio estimator is often compared and contrasted with other estimators. (See, for example, Falorsi, P.D. and Russo, A. (1999).)

Important Properties:

There are various reasons that the CRE makes good common sense: It appears robust in practice, as discussed at the end of this section; it is written in terms of only $\sum x_i$ and

$\sum y_i$, yet results in reasonable heteroscedasticity; it is intuitively obvious that $\bar{y}/\bar{x} \approx \bar{Y}/\bar{X}$ (the ratio of means of the y_i and x_i values from the sample is approximately equal to the corresponding ratio based on the entire population) and thus, a good approximation for $\sum_{i=1}^{N} y_i$ is $[\sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i] \sum_{i=1}^{N} x_i$, because, in practice, it appears that any bias in $\bar{y}$ is likely to correspond to a similar bias in $\bar{x}$. Further, a zero intercept makes sense if *y* should be zero when *x*=0; and the CRE even has the property that OLS claims regarding the sum of estimated residuals being zero. The sum of the residuals is zero for OLS regression, γ being zero and the intercept not being fixed at zero, but it is also zero for $\gamma = 0.5$ if the intercept *is* set to zero (the CRE). Further, Brewer (2002), page 43, Theorem 3.5.1 states that the CRE is "Cochran consistent" (i.e., when n=N, the bias in the estimate becomes zero) for simple random sampling without replacement, "… but the 'average of ratios' [$\gamma = 1.0$] estimator, in general, is not." We will see that the CRE has good properties under design-based and model-based sampling, and for imputation.

The classical ratio estimator (CRE) is a weighted least squares (WLS) estimator, and is discussed in Brewer (2002), pages 108-111, and elsewhere. It is compared to the simple regression estimator, an ordinary least squares (OLS) method. On page 108 of Brewer (2002), we have "It is our considered opinion, however, that the circumstances in which simple regression estimation is preferable to classical ratio estimation do not occur very often." On pages 109-110, the idea that an intercept term makes a model more flexible is questioned. Further, the following is stated:

> It is more often the case than not, in survey sampling, that the most appropriate supplementary variable is close to being proportional to its corresponding survey variable, and that their natural relationship or line of best fit is a straight line through the origin. If the range of the supplementary variable is limited (and it often is limited by the process of stratification on size) then the inclusion of an intercept term permits the estimated relationship to stray well away from the origin, with a consequent loss of efficiency.

The Role of Heteroscedasticity:

On page 111 of Brewer (2002), a discussion of reasonable values for the coefficient of heteroscedasticity, γ, for establishment surveys includes the CRE, with $\gamma = 0.5$. On that page it also mistakenly says that for the CRE, the mean of the estimated residuals is not generally equal to zero, although the sum of the expected residuals is zero for this and any best linear unbiased estimator (BLUE). A proof Ken Brewer supplied more recently

stipulates that the estimated residuals for the CRE do always sum to zero. This proof is given in a section below.

Note then that although many may think it a redeeming quality that the estimated residuals in OLS regression always sum to zero, this is also true of the CRE, and the assumed variance structure is more reasonable. In fact, especially for smaller numbers of observations, it is often not practical to rely on estimations of the degree of heteroscedasticity (Knaub (2002)), although methods of estimation are available: Carroll and Ruppert (1988), Knaub (1997), Brewer (1963). Sweet and Sigman (1995), Section 2.6, mentions references to such procedures. "Portability" of the coefficient of heteroscedasticity from one dataset to the next in a given series is problematic. (See Brewer, Foreman, Mellor, and Trewin (1977), and also Knaub (1995), where comments are made on the usefulness of employing $\gamma = 0.5$, the CRE, especially at more aggregate data collection levels. Note the first paragraph on page 704. Also note the first paragraph on page 701.)

More on the CRE:

Further positive remarks on the CRE are found in Cochran (1977). On page 158, Theorem 6.3 says "Under [a] model [where the y-x relationship is linear through the origin, and the variance of the y's is proportional to the x-values] the ratio estimator $\hat{Y}_R = X\,\bar{y}/\bar{x}$ is a best linear unbiased estimator for any sample, random or not, selected solely according to the values of the x_i." On that page, Cochran gave credit to Brewer (1963) and Royall (1970). On pages 159 and 160, Cochran (1977), we see "When we are trying to decide what kind of estimate to use, a graph in which the sample values of y_i are plotted against those of x_i is helpful. If this graph shows a straight line relation passing through the origin and if the variance of the points y_i about the line seems roughly proportional to x_i, the ratio estimator will be hard to beat." "Hard to beat" seems the operative phrase. The CRE at least appears impossible "to beat" consistently. Even when slight improvements may be made, the question is "How stable are they?" For instance, when producing monthly estimates, can one count on any estimator to more consistently produce good estimates? Experience of the author at the US Energy Information Administration seems to confirm that a ratio estimator really is "hard to beat." It will not always produce the best results, but it seldom performs badly.

On page 160, Cochran (1977), there is reference to a somewhat extreme case, (Jessen, et al. (1947)). There the situation is such that $\gamma = 1.0$. It is obvious in that case that the heteroscedasticity is unusually extreme. The CRE appears to cover a far wider range of more natural occurrences.

In Knaub (2001), it is mentioned that using γ = 0.5 seemed useful. On page 10 of that reference, test data were exhibited graphically that showed a smaller range in differences between observed and estimated totals for that test set when γ = 0.5 was used than when γ = 0.7 was used with adjustments to reduce bias.

In Chapter 15, "Heteroskedastic Errors," Griffiths, Hill and Judge (1993), they introduce the reader to weighted least squares regression in an economic setting. On page 485 they say "Given that our inspection of the least squares residuals suggests that the error variance increases as income increases, a reasonable model for the variance relationship is

$$\text{var}(e_i) = \sigma_i^2 = \sigma^2 x_i$$

They caution that there should not only be an indication of heteroscedasticity in the data, but that there should be economic reasoning that would explain it. This particular variance structure is that of the CRE, if the intercept is set to zero. It seems to have been found useful by a variety of researchers. The logical use of the CRE in an economic setting also appears supported by its appearance in Maddala (1977), pages 259-261.

Lohr (1999), pages 81-82 also discusses the classical ratio estimator in the context of survey data. Her treatment of totals, like the Brewer proof below, relies on the relationship $\hat{\beta} = \bar{y}/\bar{x}$. She refers to the resulting estimation of a total from a survey as a "natural estimate."

The terms "reasonable model," and "natural estimate" seem very appropriate with regard to the classical ratio estimator. It seems a logical approach to a variety of problems.

Asides on variance:
(1) From Knaub (2004), page 12, and Knaub (1992), pages 776-777, with regard to one regressor for the CRE, and with regard to the unobserved values, a sum of those regressor data is sufficient to aid in estimating the variance of the estimated total. Thus, if regressor data are from a source such that it is difficult or impossible to make a one-to-one match between all regressor values and the members of the population of interest, we may still estimate a variance for the total from the CRE.
(2) Royall and Cumberland (1981) studied variance for the ratio estimator to make an improvement, but Knaub (1992) found the added complexity in adjusted variance estimation may not be justified. Knaub (2002) suggests the desirability of simplicity.

A final introductory note concerns robust regression. Such techniques are designed to reduce the influence of outliers. This, however, may often assume more information is available than may actually be present. WLS regression may already reduce the outlier problem since the influence of the largest observations is reduced. Most importantly, any responses that 'fail' edits should be pursued with the respondents. The CRE is generally enough. Other adjustments may just bias results. (Note, however, that when WLS regression is used, an outlier near the origin may greatly inflate variance estimates for an

estimated total.) Chapter 14 in Brewer (2002) discusses extreme values. Pragmatism has to take a leading role in their treatment. Sarndal and Lundstrom (2005) stresses reduction of bias due to nonresponse by attention to stratification. Knaub (1999) also stresses attention to grouping data under models. Thus, whether design-based or model-based sampling and estimation, the CRE can be improved by attention to how the data are grouped for estimation purposes.

Zero Sum of Estimated Residuals:

From Knaub (2004), page 12, "$\sum y_i = \sum \overset{*}{y}_i$ when $\gamma = 0.5$, for the model-based ratio estimator," the intercept being set at zero. Therefore, under the CRE, the estimated residuals must always sum to zero. (It happens that this can also be a good check for a computer programming error.)

A simple proof due to Brewer (personal correspondence), clearly shows that under the CRE, estimated residuals do always sum to zero:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left\{ y_i - \hat{\beta} x_i \right\}, \text{ where } \hat{\beta} = \sum_{i=1}^{n} y_i \Big/ \sum_{i=1}^{n} x_i \text{, hence}$$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left\{ y_i - \left(\sum_{j=1}^{n} y_j \Big/ \sum_{j=1}^{n} x_j \right) x_i \right\} = \sum_{i=1}^{n} y_i - \left(\sum_{i=1}^{n} y_i \Big/ \sum_{i=1}^{n} x_i \right) \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i = 0.$$

Although this is a 'special' case of weighted least squares, it is actually quite generally applicable in many situations, as discussed above.

An interesting relationship is to be seen between this proof and a property noted in Fox (1997). John Fox shows that any line through the means of the variables, $(\bar{x}, \bar{y})$, has $\sum_{i=1}^{n} e_i = 0.$

Consider an intercept (or not), and one (or more) regressors, with residuals that may or may not be heteroscedastic:

$$y_i = \hat{\alpha} + \hat{\beta} x_i + e_i$$

From Fox, if one solution to $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ is $\bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}$, then $\sum_{i=1}^{n} e_i = 0.$

Therefore, $\hat{\beta} = (\bar{y} - \hat{\alpha}) / \bar{x}$ when $\sum_{i=1}^{n} e_i = 0.$

This is consistent with the above proof that $\sum_{i=1}^{n} e_i = 0$ for the CRE, since $\hat{\beta} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i$ means that $\hat{\beta} = \bar{y} / \bar{x}$, and that works above if $\hat{\alpha} = 0$. (Further, a section on the axiom by Fox, and a corollary to it, appears below.)

Note also that in Brewer (2002), page 114, for OLS, "… that if two continuously distributed variables, x and y, both having positive means, are positively correlated, and if the natural relationship between them passes through the origin, then the regression of y on x has a positive intercept on the [errata: should be y-axis] and the regression of x on y has a positive intercept on the [x-axis]… ." Each line, in such a case, passes through $(\bar{x}, \bar{y})$ according to Fox (1997). When the intercept is set at the origin, then the WLS case with variance proportional to x results, and we have the CRE. Webster and Oliver (1990) show a related figure on page 106.

Conditions Under Which Estimated Residuals Sum to Zero:

Circa August 2004, inquiries were made on the internet listservs for the Government Section and for the Survey Research Methods Section of the American Statistical Association, requesting references regarding zero sums of estimated residuals, and among the responses were two that will be mentioned here: First, Alan Dorfman referenced a lemma in Valliant, Dorfman and Royall (2000), Lemma 4.2.1, page 97. Later, Phil Kott noted pages 231-232 of Sarndal, Swensson and Wretman (1992). These references gave conditions under which estimated residuals will sum to zero. Both OLS regression and the CRE satisfy those conditions.

First, considering pages 231 and 232, Sarndal, Swensson and Wretman (1992), equation 6.5.7 and Example 6.5.1, let us examine possibilities involving a single regressor**:**

Let $y_i = \hat{\alpha} + \hat{\beta} x_i + e_i$, then

$$\sum_{i=1}^{n} y_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} e_i \, .$$

Can we show that $\sigma_i^2 \propto x_i$ means $\sum_{i=1}^{n} e_i = 0$ and $\hat{\alpha} = 0$?

Because $\sigma_i^2 \neq \sigma^2$, we will use $e_i = e_{0i} x_i^{\gamma}$.

In that case, $\sigma_i^2 \propto x_i$ means that $\gamma = 0.5$.

If $\gamma = 0.5$, then $\hat{\beta} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i$, so

$$\sum_{i=1}^{n} y_i = n\hat{\alpha} + \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} e_i \Rightarrow \sum_{i=1}^{n} e_i = -n\hat{\alpha}.$$

So, if $\gamma = 0.5$, $\sum_{i=1}^{n} e_i = 0$ only if $\hat{\alpha} = 0$.

So, yes, $\sigma_i^2 \propto x_i \Rightarrow \sum_{i=1}^{n} e_i = 0$ if $\hat{\alpha} = 0$.

Example 6.5.1 in Sarndal, Swensson and Wretman (1991) shows three "examples" of "variance structures" that satisfy the condition shown there, such that the sum of the estimated residuals is zero. The condition states that the variance must be proportional to a regressor or linear combination of regressors. Those are clearly represented by the second and third "examples" on page 232. An additional choice is the OLS case. With OLS regression, variance is proportional to a constant because there is an intercept term, and variance is considered constant. Thus OLS regression technically meets the condition, but with an unrealistic variance structure assumed for survey data. The variance is proportional to a regressor in the case of the CRE. That was the second part of Example 6.5.1. In the case of multiple regression, the variance may be proportional to any of the regressors, or any linear combination of them. That was the third part of Example 6.5.1 in Sarndal, Swensson and Wretman (1992). In Knaub (2003), pages 3-5, "Regression Weights," it was independently suggested that a linear combination of regressors would make a good "size" estimate "z" for use in regression weights of the form $w_i = z_i^{-2\gamma}$. With $\gamma = 0.5$, this satisfies the condition leading to $\sum_{i=1}^{n} e_i = 0$. One good possibility for z might be $z_i = \tilde{y}_i$, where $\tilde{y}_i$ would be a preliminary estimate of y_i. For $\gamma = 0.5$, this would be a multiple regression extension of the CRE, and once again, the intercept must be at the origin.

Alternative Variance Structures and Other Considerations:

In Maddala (1977), he uses $\beta^* = \bar{y}/\bar{x}$ to describe the classical ratio estimator, although, of the references for this article, apparently only Brewer (2002) uses the name "classical ratio estimator." Maddala uses an asterisk rather than a "hat" to designate estimates of the coefficient β, when WLS regression is used. (Note, however, that OLS regression is just a special case of WLS regression.) On page 261 of Maddala (1977), he mentions work by Prais and Houthakker, published in 1955, where they apparently use the square of $\tilde{y}_i$, or a closer estimate of y_i. From Maddala (1977), page 261, they have σ_i^2 "… proportional to the square of the regression function; i.e., $\sigma_i^2 = \sigma^2(\alpha + \beta x_i)^2$." Thus, this should be more like the situation found by Jessen, et al. (1947), where $\gamma = 1$. If instead they had used $\sigma_i^2 = \sigma^2(\alpha + \beta x_i)$, then that would have agreed better with Knaub (2003), although the latter dealt with multiple regression and no intercept. If we use WLS, the intercept term should not be present as part of the condition for having $\sum_{i=1}^{n} e_i = 0$.

Note that in some cases the MSE under OLS may be smaller, but when OLS is obviously incorrect, a smaller MSE does not excuse its use.

When determining what data to place under a given regression model, in the case of a single regressor, one may compare regression coefficients for groups that are candidates for being combined. This requires attention to the standard deviations of those coefficients. For multiple regression, a single regressor such as used in the regression weights ($\tilde{y}$ or z), a linear combination of regressors, could be used in this case to give us a common frame of reference, for this purpose only.

Fox's Axiom on Sum of Estimated Residuals and a Corollary

In Fox (1997), pages 86 and 87, we see a development which leads to the following axiom:

Fox's Axiom:

If a regression line with one regressor goes through the point $(\bar{x}, \bar{y})$, the means of the variables, then the sum of the estimated residuals is zero.

Proof:

From page 86 we consider $y_i = \hat{\alpha} + \hat{\beta} x_i + e_i$. On page 87, consider what happens when the point $(\bar{x}, \bar{y})$ is on the regression line: $\bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}$

If we subtract the second equation from the first, Fox proceeded as follows:

$$y_i - \bar{y} = \hat{\beta}(x_i - \bar{x}) + e_i$$

$$e_i = y_i - \bar{y} - \hat{\beta}(x_i - \bar{x})$$

Summing,

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \bar{y}) - \hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x}) = 0 - (\hat{\beta})0 = 0$$

……..….

What of the converse? That proves to be true also.

Corollary to Fox's Axiom:

If the estimated residuals sum to zero, then the point $(\bar{x}, \bar{y})$ must lie on the regression line.

Proof:

Starting with $y_i = \hat{\alpha} + \hat{\beta} x_i + e_i$, we have the following:

$$e_i = y_i - \hat{\alpha} - \hat{\beta} x_i$$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)$$

If $\sum_{i=1}^{n} e_i = 0$, then $\sum_{i=1}^{n} y_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} x_i$.

Thus $\bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}$.

Therefore, if the estimated residuals sum to zero, then the point $(\bar{x}, \bar{y})$ must lie on the regression line.

Combining Fox's axiom (Fox (1997)) and the corollary here, we have the following:

The sum of estimated residuals for a linear regression with one regressor is equal to zero if and only if the regression goes through the point $(\bar{x}, \bar{y})$.

When the sum of residuals is zero, the sum of observations, and the sum of the predictions that would replace them, would be equal. That is obviously desirable from the standpoint of accuracy of predictions, conditional upon the data collected, but it also can be quite helpful when "debugging" software, as it provides a means for checking some results.

Zero Sum of Estimated Residuals and Estimated Weighted Residuals:

First, for intercept fixed at 0, and $\gamma = 0.5$, *i.e.*, the classical ratio estimator (CRE), we have $\sum_{i=1}^{n} e_i = 0$.

Second, for intercept fixed at 0, and $\gamma = 1.0$, we have $\sum_{i=1}^{n} e_{oi} = 0$, where $e_i = z_i^{\gamma} e_{0i}$, and for one regressor, generally $z = x$.

In the first case we have the Ken Brewer proof:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left\{ y_i - \hat{\beta} x_i \right\} = \sum_{i=1}^{n} y_i - \hat{\beta} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i - \left(\sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i \right) \sum_{i=1}^{n} x_i$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i = 0.$$

In the latter case, we have the following:

$$y_i - \hat{\beta} x_i = x_i e_{0i}.$$

Therefore, $\dfrac{y_i - \hat{\beta} x_i}{x_i} = e_{0i} = \dfrac{y_i}{x_i} - \hat{\beta}(\gamma = 1)$, and thus

$$\sum_{i=1}^{n} e_{oi} = \sum_{i=1}^{n} \frac{y_i}{x_i} - \sum_{i=1}^{n} \hat{\beta}(\gamma = 1) = 0$$

This proof has nothing to do with the way data were actually generated. (See Table 1 and Table 2.) However, if it is reasonable that y=0 when x=0, and the establishment survey data obtained normally indicate $0.5 \le \gamma \le 1.0$, and we desire that $\sum_{i=1}^{n} e_{0i} \cong 0$ with a reasonable heteroscedastic model, and $\sum_{i=1}^{n} e_i \cong 0$ is always desirable, then models with the intercept set to zero and γ set somewhere in the range $0.5 \le \gamma \le 1.0$ should perform well. (See Brewer (2002), page 111.) Knaub (1993), and Knaub (2001) discuss the robust nature of using $\gamma = 0.5$. Further, Brewer (2002), page 134, discusses the advantage of picking γ toward the lower range of $0.5 \le \gamma \le 1.0$, to avoid unacceptably low design-based sample weights due to calibration, even though the data usually support γ closer to the middle of that range.

Draper and Smith (1998), page 222, shows a transformation from a heteroscedastic to a homoscedastic model, which ensures that the expected value of the transformed residuals will be zero. Nancy Kirkendall pointed out that these transformed residuals are what are output in SAS PROC REG. Note that the transformation of residuals here is also what was used in Knaub (1993) and Knaub (1997), page 2, to estimate the coefficient of heteroscedasticity in a manner that often performs better than the iterated reweighted least squares (IRLS) method. In Knaub (1995), Appendix I, pages 704-705, a question is raised of a better format for regression weights. In Steel and Fay (1995), page 375, a suggestion is offered. However, they also show the very simple format used here, except for the intercept term, perhaps first shown in Cochran (1953). In general, that format performs well. Further simplifying by letting $\gamma = 0.5$ accounts well for a natural degree of heteroscedasticity, while avoiding overspecification of parameters, which seems to be common with the available data in many if not most practical circumstances, particularly with survey data. (See also Sweet and Sigman (1995) regarding use of $\gamma = 0.5$.)

CRE - Relationships to Other Estimators:

It is known that under the same model needed for the CRE,

$$y_i = \hat{\beta} x_i + e_i \text{ with } \sigma_i^2 \propto x_i,$$

the generalized regression estimator (GREG) for the population total becomes equal to the standard ratio estimator. This is shown in Lohr (1999), page 373, Example 11.9, and in Estevao, Hidiroglou and Sarndal (1995). On page 42 of Sarndal and Lundstrom

(2005), it is shown that the simple regression estimator and the ratio estimator are both special cases of the GREG. Both regression weights and design weights are included in the formulations.

As another example of the relationship of the CRE to other statistics, consider the relative standard error with regard to a superpopulation (the RSESP, Knaub (2004)). The RSESP using $\gamma = 0.5$ is the subject of current study for the purpose of measuring total survey error. Another measure of total survey error could be the root mean square error (RMSE), which may be seen as a component of the RSESP, ignoring the contribution to uncertainty due to estimating the regression coefficient(s). Consider the "exact" formula for the model-based RSESP with one regressor and a zero-intercept. For this case, the general form of the model used will be as follows:

$y_i = \sum_{h=1}^{H} b_h x_{ih} + w_{ih}^{-0.5} e_{0_{ih}}$, where H is the number of estimation groups (strata). Note that $e_{ih} = w_{ih}^{-0.5} e_{0_{ih}}$, so $w_{ih}^{-0.5}$ and $e_{0_{ih}}$ are the 'nonrandom' and 'random' factors of the residual (Knaub (1995)), respectively.

Let N be the population size, taken from the superpopulation of infinite size from which it was generated, and let n be the size of the sample collected.

The "exact" formula of the estimated, model-based relative standard error, with respect to the superpopulation, the model-based RSESP, is as follows**:**

$$RSESP = \left(\hat{V}^{*}_{L,SP}(\hat{T}^{*} - T)\right)^{0.5} / \hat{T}^{*}$$

with $$\hat{V}^{*}_{L,SP}(\hat{T}^{*} - T) = \sum_{h=1}^{H} \left[\sum_{i=1}^{N_h} \frac{\sigma^{*2}_{e_0 h}}{w_{ih}} + \left(\sum_{i=1}^{N_h} x_{ih} \right)^2 \hat{V}^{*}(b_h) \right]$$

where $\sigma^{*2}_{e_0 h} = \sum_{i=1}^{n'_h} \frac{e^2_{0_{ih}}}{n'_h - 1}$, n'_h is the sample size in stratum h, without add-ons, and H is the number of estimation groups (*i.e.*, strata).

"Add-ons" (Knaub (2002)) constitute an additional uncertainty that is not measured. As formulated here, using the classical ratio estimate, we are assuming that the estimated standard error of the random factor of the residuals, when multiplied by the nonrandom factor, $w_{ih}^{-0.5} = x_{ih}^{\gamma} = x_{ih}^{0.5}$, is representative of the error in each case, including the add-ons. However, in reality, the uncertainty due to add-ons is not the same, by definition, and could be quite different.

The key to the relationship between the RMSE and the RSESP is in the formula

$$\overset{*}{V}_{L,SP}(\overset{*}{T}-T) = \sum_{h=1}^{H}\left[\sum_{i=1}^{N_h}\frac{\overset{*}{\sigma}^{2}_{e_0 h}}{w_{ih}} + \left(\sum_{i=1}^{N_h} x_{ih}\right)^2 \overset{*}{V}(b_h)\right].$$

Notice that in $\left[\sum_{i=1}^{N_h}\frac{\overset{*}{\sigma}^{2}_{e_0 h}}{w_{ih}} + \left(\sum_{i=1}^{N_h} x_{ih}\right)^2 \overset{*}{V}(b_h)\right]$, there is a sum of squared errors component and a component due to the estimation of the coefficient of the regressor. Therefore, because the estimated 'MSE' here is conditional upon data collected from an infinite population, and estimated residuals from a model, it is not as encompassing as an MSE that is the traditional sum of variance and bias. Even in the case of a finite population, the estimated residuals make this estimated MSE only a part of the estimated variance of the estimated total.

Conclusion:

The CRE is very simple, versatile and useful, especially for survey statistics. It may be used as part of a larger analysis, and/or as a stable means of producing high quality data for publication in a timely manner.

Table 1

Two Models Applied to Data Generated With γ=1 and Nonzero y-intercept
(See Knaub (1997), Figure 2.0 and Figure 2.1.)

γ=1.0 in generated data, y

CRE: $y_i' = [b(\gamma=0.5)]x_i + (\sqrt{x_i})e0_i$

γ=0.5 true intercept=100

X	y	y'	CRE: **e0i**	CRE: **ei**=(√xi)e0i	
100	80	180	4.000000	40.0000	
100	120	220	8.000000	80.0000	
200	160	260	-1.414214	-20.0000	
200	240	340	4.242641	60.0000	
300	240	340	-4.618802	-80.0000	
300	360	460	2.309401	40.0000	
400	320	420	-7.000000	-140.0000	
400	480	580	1.000000	20.0000	
2000	**2000**	**2800**	**6.519026**	**0.0000**	**<- SUMs**

b(γ=0.5) = 1.4000

Alternate Model: $y_i' = [b(\gamma=1.0)]x_i + (x_i)e0_i$

γ=1.0 true Intercept=100

X	y	y'	**e0i**	**ei**=(xi)e0i	
100	80	180	0.279167	27.9167	
100	120	220	0.679167	67.9167	
200	160	260	-0.220833	-44.1667	
200	240	340	0.179167	35.8333	
300	240	340	-0.387500	-116.2500	
300	360	460	0.012500	3.7500	
400	320	420	-0.470833	-188.3333	
400	480	580	-0.070833	-28.3333	
2000	**2000**	**2800**	**0.000000**	**-241.6667**	**<- SUMs**

b(γ=1.0) = 1.5208

Table 2

Two Models Applied to Data Generated With γ=0.5 and Nonzero y-intercept
(See Knaub (1997), Figure 2.0 and Figure 2.1.)

γ=0.5 in generated data, y

CRE: $y_i' = [b(\gamma=0.5)]x_i + (\sqrt{x_i})e0_i$

γ=0.5 true Intercept=1000

x	y	y'	CRE: **e0i**	CRE: **ei**=($\sqrt{x_i}$)e0i	
1600	1290.2	2290.2	-3.797632	-151.9053	
1600	1909.8	2909.8	11.692368	467.6947	
1800	1471.4	2471.4	-6.504638	-275.9684	
1800	2128.6	3128.6	8.985714	381.2316	
2000	1653.6	2653.6	-8.922617	-399.0316	
2000	2346.4	3346.4	6.568862	293.7684	
2200	1836.7	2836.7	-11.111909	-521.1947	
2200	2563.3	3563.3	4.379255	205.4053	
15200	**15200**	**23200**	**1.289403**	**0.0000**	**<- SUMs**

b(γ=0.5) = 1.5263

Alternate Model: $y_i' = [b(\gamma=1.0)]x_i + (x_i)e0_i$

γ=1.0

true intercept=1000

x	y	y'	**e0i**	**ei**=(x_i)e0i	
1600	1290.2	2290.2	-0.102400	-163.8404	
1600	1909.8	2909.8	0.284850	455.7596	
1800	1471.4	2471.4	-0.160775	-289.3955	
1800	2128.6	3128.6	0.204336	367.8045	
2000	1653.6	2653.6	-0.206975	-413.9505	
2000	2346.4	3346.4	0.139425	278.8495	
2200	1836.7	2836.7	-0.244366	-537.6056	
2200	2563.3	3563.3	0.085907	188.9944	
15200	**15200**	**23200**	**0.000000**	**-113.3838**	**<- SUMs**

b(γ=1.0) = 1.5338

Acknowledgement:

I thank K.R.W. Brewer for his generosity in providing several helpful suggestions.

References:

Herv´e Abdi (2003), "Least Squares," Lewis-Beck M., Bryman, A., Futing T. (Eds.), Encyclopedia of Social Sciences Research Methods. Thousand Oaks (CA): Sage. (http://www.utdallas.edu/~herve/Abdi-LeastSquares-pretty.pdf)

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, KRW (2002), Combining survey sampling inferences: The weighing of Basu's elephants, Arnold: London and Oxford University Press.

Brewer, K.R.W., Foreman, E.K., Mellor, R.W., and Trewin, D.J. (1977), "Use of Experimental Design and Population Modelling in Survey Sampling," Bulletin of the International Statistical Institute, 47, pp. 173-190.

Carlson, S.R., L.G. Coggins, and C.O. Swanton. 1998. "A Simple Stratified Design for Mark-Recapture Estimation of Salmon Smolt Abundance." Alaska Fishery Research Bulletin 5: 88-102.

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

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

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

Cochran, W.G. (1978), "Laplace's Ratio Estimator," Contributions to Survey Sampling and Applied Statistics, ed. HA David. New York: Academic Press., 3-10.

Draper, N., and H. Smith (1998), Applied Regression Analysis. 3rd ed., New York: Wiley.

Estevao, V., Hidiroglou, M.A., and Sarndal, C.E. (1995). "Requirements on a Generalized Estimation System at Statistics Canada," Journal of Official Statistics, pages 181-204, Statistics Sweden. (http://www.jos.nu)

Falorsi, P.D. and Russo, A. (1999). "A Conditional Analysis of some Small Area Estimators in Two Stage Sampling." Journal of Official Statistics, pages 537-550, Statistics Sweden. (http://www.jos.nu)

Fox, J. (1997), Applied Regression Analysis, Linear Models, and Related Methods, Sage Publications.

Griffiths, W.E., Hill, R.C., and Judge, G.G., Learning and Practicing Econometrics, Wiley 1993.

Jessen, Raymond J., *et.al*., "On a Population Sample for Greece," Journal of the Statistical Association, Vol. 42, September, 1947.

Knaub, J.R., Jr. (1992), "More Model Sampling and Analyses Applied to Electric Power Data," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 876-881.

Knaub, J.R., Jr. (1993), "Alternative to the Iterated Reweighted Least Squares Method: Apparent Heteroscedasticity and Linear Regression Model Sampling," Proceedings of the International Conference on Establishment Surveys, American Statistical Association, pp. 520-525.

Knaub, J.R., Jr. (1995), "A New Look at 'Portability' for Survey Model Sampling and Imputation," Proceedings of the Section on Survey Research Methods, Vol. II, American Statistical Association, pp. 701-705.

Knaub, J.R., Jr. (1997), "Weighting in Regression for Use in Survey Methodology," InterStat, April 1997, http://interstat.statjournals.net/. (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 1997.)

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.)

Knaub, J.R., Jr. (2002), "Practical Methods for Electric Power Survey Data," InterStat, July 2002, http://interstat.statjournals.net/. (Note another version in ASA Survey Research Methods Section proceedings, 2002.)

Knaub, J.R., Jr. (2003), "Applied Multiple Regression for Surveys with Regressors of Changing Relevance: Fuel Switching by Electric Power Producers," InterStat, May 2003, http://interstat.statjournals.net/. (Note another version in ASA Survey Research Methods Section proceedings, 2003.)

Knaub, J.R., Jr. (2004), "Modeling Superpopulation Variance: Its Relationship to Total Survey Error," InterStat, August 2004, http://interstat.statjournals.net/. (Note another version in ASA Survey Research Methods Section proceedings, 2004.)

Liu, Y., Batcher, M. and Scheuren, F. (2005), "Efficient Sampling Design in Audut Data," Journal of Data Science," v.3, no.3, pages 213-222.
(http://www.sinica.edu.tw/~jds/)

Lohr, S.L. (1999), Sampling: Design and Analysis, Duxbury Press.

Maddala G S. (1977), Econometrics. New York: McGraw-Hill.

Royall, R.M. (1970), "On Finite Population Sampling Theory Under Certain Linear Regression Models," Biometrika, 57, pp. 377-387.

Royall, R.M. and Cumberland, W.G. (1981), "An Empirical Study of the Ratio Estimator and Estimators of its Variance," Journal of the American Statistical Association, 76, pp. 66-88.

Sarndal, C.-E. and Lundstrom, S. (2005), Estimation in Surveys with Nonresponse, John Wiley& Sons, Ltd.

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

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

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

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

Webster, R., and Oliver, M.A. (1990), Statistical Methods in Soil and Land Resource Survey, Oxford University Press.