RELATIVE STANDARD ERROR FOR A RATIO OF VARIABLES AT AN AGGREGATE LEVEL UNDER MODEL SAMPLING

James R. Knaub, Jr., Energy Information Administration
U.S. Department of Energy, EI-521, Washington, D.C. 20585

Key Words: Covariance, Model weights

Abstract:
One of the surveys conducted by the Energy Information Administration (EIA), Form EIA-826, "Monthly Electric Utility Sales and Revenue Report with State Distributions," collects sales and associated revenue, by month, on a (model) sample of U.S. electric utilities. A model is used to estimate sales, revenue, and revenue per kilowatthour. The Form EIA-826 data are regressed on similar data from the Form EIA-861, "Annual Electric Utility Report," which is an annual census of U.S. electric utilities. (Investigation has shown that only the largest utilities need to be sampled, thus reducing the reporting burden on the smaller utilities.) There is a correlation between sales and revenue, thus for revenue per kilowatthour, the relative standard error (RSE), often referred to as the coefficient of variation (CV), is subject to a covariance (COV) term. A COV term for totals (instead of model coefficients) involved in model sampling/estimation is not commonly found in the literature. One developed by Prof. Poduri S.R.S. Rao (University of Rochester, and EIA) is used and presented here (see Rao (1992)). These estimates are all done at the State level. The next step is to aggregate results to groups of States described by the Census Bureau (i.e., Census divisions) and to the national level. To estimate CV's at aggregate levels for revenue per kilowatthour, even though the States' data are all considered to be independent, is a little easier to do incorrectly than one might think. My first method for accomplishing this did not yield reasonable results. (Thanks to Stephen Calopedis, EIA, for discovering this.) My final method did yield reasonable results, and I later found support for it in Hansen, Hurwitz and Madow (1953).

Introduction:
From Royall (1970), for sales or revenue, for any retail sector (residential, commercial, etc.) at the State level, if we let x represent an observation from the Form EIA-861, let y represent an observation from the Form EIA-826, and let $\hat{y}$ represent an estimated value for data not collected, then $y_i = bx_i + x_i^{\gamma} e_{o_i}$, $\hat{y}_j = \hat{b}x_j$,

$$\hat{b}(\gamma) = \left[\sum_{k=1}^{n} x_k^{1-2\gamma} y_k\right] / \left[\sum_{k=1}^{n} x_k^{2-2\gamma}\right].$$

Brewer (1993) succinctly writes the estimate of the total for this best linear unbiased estimator, $\hat{T}_{BLU}(y)$, as

$$\hat{T}_{BLU}(y) = \sum_{j\in s} y_j + \frac{\sum_{j\in s} y_j x_j a_j^{-2}}{\sum_{j\in s} x_j^2 a_j^{-2}} \left(\sum_{j=1}^{N} x_j - \sum_{j\in s} x_j \right),$$

where $a_j = x_j^{\gamma}$ makes this consistent with the above. Electric power data often indicate that $0.5 < \gamma \leq 1$, but it may be more robust to use $\gamma = ½$. (See Knaub (1992) and Knaub (1993).) Here there is an average of only, approximately, n = 6 observations at the State level. Generally, the majority of the State's sales and revenue for a given retail sector are represented by 2 or 3 of these utilities. In Knaub (1993) it can be seen that such a situation may cause some disturbance in the model, which seems to be handled well by using $\gamma = ½$. Further, weights other than $x^{-2\gamma}$ could be used. In Rao (1992), weights are referred to as "w." For this application, w=1/x is used, resulting in the familiar ratio estimate, which is equivalent to the use of $\gamma = ½$ above. V_D, as found in Royall and Cumberland (1978 and 1981), is employed accordingly. (Note that Brewer (1993) goes on to show weights with perhaps better properties for many other uses.)

The following formula for covariance associated with V_D (COV_D, or, in Prof. Rao's notation, V_{12}) is due to work by Prof. Poduri S.R.S. Rao of the University of Rochester and the Energy Information Administration (EIA) found in Rao (1992), with correction by Dr. Nancy Kirkendall (EIA), also based on Rao (1992). This formula is

$$V_{12} = X_1' X_2' \frac{\sum^{n} w_{1i} x_{1i} w_{2i} x_{2i} v_{12i}}{\left(\sum^{n} w_{1i} x_{1i}^2\right)\left(\sum^{n} w_{2i} x_{2i}^2\right)} + \sum' v_{12i},$$

where a prime sign indicates a summation over the N-n members of the population not sampled, and

$$\Sigma'\hat{v}_{12i} = \left[\left(\Sigma'\frac{1}{w_{1i}^{1/2}w_{2i}^{1/2}}\right) / \left(\sum^{n}\frac{1}{w_{1i}^{1/2}w_{2i}^{1/2}}\right)\right]\sum^{n}\hat{v}_{12i},$$

$$\hat{v}_{12i} = \frac{e_{1i}e_{2i}}{1-k_{12i}},\ e_{1i} = y_{1i}-b_1x_{1i},\ e_{2i} = y_{2i}-b_2x_{2i}$$

and

$$k_{12i} =$$

$$\frac{w_{1i}x_{1i}^{2}}{\sum^{n}w_{1j}x_{1j}^{2}} + \frac{w_{2i}x_{2i}^{2}}{\sum^{n}w_{2j}x_{2j}^{2}} - \frac{w_{1i}^{1/2}w_{2i}^{1/2}x_{1i}x_{2i}\sum^{n}w_{1j}^{1/2}w_{2j}^{1/2}x_{1j}x_{2j}}{\left(\sum^{n}w_{1j}x_{1j}^{2}\right)\left(\sum^{n}w_{2j}x_{2j}^{2}\right)}$$

Now, letting $w_{1i}=x_{1i}^{-2\gamma}$ and $w_{2i}=x_{2i}^{-2\gamma}$, we can get estimates, $\hat{b}_1(\gamma)$ and $\hat{b}_2(\gamma)$, corresponding to the formula for $\hat{b}(\gamma)$ shown at the beginning of this section. (Note that, in general, the nonrandom component of error is $w_i^{-1/2}$ in $y_i=bx_i+w_i^{-1/2}e_{o_i}$, and the random component of error is e_{o_i}. However, commonly, $w_i^{-1/2} = x_i^{\gamma}$ has been the case used. Knaub (1993) and Knaub (1994) claim $\gamma = 1/2$ appears to compensate for minor model failure. Knaub (1994) showed that the best estimate of γ appears to normally depend on the range of x, and therefore, this is a commonly occurring violation of the model. This can account, for example, for an overall γ estimate of 0.8, when 0.5 may result in more accuracy when estimating the total!)

Now, for $COV_D(\gamma=1/2)$ corresponding to $V_D(\gamma=1/2)$, we take V_{12} and let $w_{1i} = x_{1i}^{-1}$ and $w_{2i} = x_{2i}^{-1}$, so

$$COV_D(\gamma=1/2) =$$

$$\left[\left(\frac{\sum^{N}x_{1i}-\sum^{n}x_{1i}}{\sum^{n}x_{1i}}\right)\left(\frac{\sum^{N}x_{2i}-\sum^{n}x_{2i}}{\sum^{n}x_{2i}}\right) + \frac{\sum^{N}x_{1i}^{1/2}x_{2i}^{1/2}-\sum^{n}x_{1i}^{1/2}x_{2i}^{1/2}}{\sum^{n}x_{1i}^{1/2}x_{2i}^{1/2}}\right]$$

$$\cdot\sum_{i}^{n}\frac{e_{1i}e_{2i}}{1-k_{12i}}$$

where

$$e_{1i} = y_{1i} - \left(\frac{\sum^{n}y_{1k}}{\sum^{n}x_{1k}}\right)x_{1i},$$

$$e_{2i} = y_{2i} - \left(\frac{\sum^{n}y_{2k}}{\sum^{n}x_{2k}}\right)x_{2i}, \text{ and}$$

$$k_{12i} = \frac{x_{1i}}{\sum^{n}x_{1j}} + \frac{x_{2i}}{\sum^{n}x_{2j}} - \frac{x_{1i}^{1/2}x_{2i}^{1/2}\sum^{n}x_{1j}^{1/2}x_{2j}^{1/2}}{\left(\sum^{n}x_{1j}\right)\left(\sum^{n}x_{2j}\right)}.$$

$COV_D\ (\gamma=1/2)$ is the covariance used when estimating State level CV's for revenue per kilowatthour for a given retail sector. This seems to be a reasonable approach for this application. However, please note that in Knaub (1992), it is shown that when comparing Royall and Cumberland's V_D to the supposedly less robust (more model dependent) V_L, V_L may look better than one would think. Also, V_L is easier to calculate, and even with modern computers, it is still possible to exceed their limitations. Further, V_L lends itself to a more direct study of a model and its fit. For these, and perhaps other reasons, the reader may also be interested in COV_L as found in Rao (1992), so it is also given here.

$$COV_L =$$

$$\left[X_1'X_2'\frac{\sum^{n}w_{1i}^{1/2}w_{2i}^{1/2}x_{1i}x_{2i}}{\left(\sum^{n}w_{1i}x_{1i}^{2}\right)\left(\sum^{n}w_{2i}x_{2i}^{2}\right)} + \Sigma'\frac{1}{w_{1i}^{1/2}w_{2i}^{1/2}}\right]$$

$$\cdot\left[\frac{1}{n-1}\sum^{n}w_{1i}^{1/2}w_{2i}^{1/2}(y_{1i}-b_1x_{1i})(y_{2i}-b_2x_{2i})\right]$$

Method at the Aggregate Level:

The problem here is finding $\hat{CV}$ for Census division and national level revenue per kilowatthour. Call such an aggregate level CV estimate here $\hat{CV}_{D_R}$. Letting $\hat{\sigma}^2_{D_{R_i}}$ be the estimated variance for State i, $\hat{T}_{N_i}$ be the estimated total revenue in State i, and $\hat{T}_{D_i}$ be the corresponding total sales estimate, it may at first appear that $\hat{CV}_{D_R} = \left[\sum_i \hat{\sigma}^2_{D_{R_i}}\right]^{1/2} \Big/ \left[\sum_i \hat{T}_{N_i} / \sum_i \hat{T}_{D_i}\right]$, but, since revenue per kilowatthour is not summed, this is incorrect! (Note that when I showed this, and my solution I eventually proposed, to another statistician, that statistician immediately derived something equivalent to the mistake above! This anecdote may be coincidental, or it may be that, in one form or another, it is easy to make a subtle mistake that could result in something equivalent to this error.)

What we really need is, letting N represent revenue and D represent sales (numerator and denominator in, for example, revenue per kilowatthour),

$$\hat{CV}_{D_R} = \left| \hat{CV}^2_{RAL} + \hat{CV}^2_{SAL} - 2\,\frac{COV_{D_{N,D}}}{\left(\sum_i \hat{T}_{N_i}\right)\left(\sum_i \hat{T}_{D_i}\right)} \right|^{1/2},$$

where "RAL" means "revenue at aggregate level," and "SAL" means "sales at aggregate level." (Also note that the "D" in CV_D and COV_D relates them to V_D, so $COV_{D_{N,D}}$ has two "D" subscripts, but they represent different interests.)

Now, write this as

$$\hat{CV}_{D_R} = \left| \frac{V_{D_N}}{\hat{T}_N^2} + \frac{V_{D_D}}{\hat{T}_D^2} - 2\,\frac{COV_{D_{N,D}}}{\hat{T}_N \hat{T}_D} \right|^{1/2}.$$

Can the aggregate level COV, namely $COV_{D_{N,D}}$, be related to the disaggregate (State) level COV_D values? It seems reasonable that

$COV_{D_{N,D}} = \sum_i COV_{D_{N,D_i}}$. That is, the aggregate level covariance should be somewhere near the sum of the covariances for each State part. In Hansen, Hurwitz and Madow (1953), pages 56-58, we see that this is exactly the case, when the State level values are uncorrelated. This is nearly specifically shown in a corollary to Theorem 12 in that book.

Conclusions:

From this paper one can see exactly how aggregate level revenue per kilowatthour CV's are calculated from a monthly sample of electric utilities which EIA then publishes in the Electric Power Monthly. Also, one can use this paper as an easy reference to apply any weighted form of V_L or V_D to an aggregate level for a ratio of variables. (Note V is found from a COV formula by letting $x_1 = x_2$ and $y_1 = y_2$.)

References:

Brewer, K.R.W. (1993), "Combining Design-Oriented and Model-Oriented Inference," to appear in the monograph for the International Conference on Establishment Surveys, Buffalo, NY, 1993, John Wiley & Sons.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G.(1953), Sample Survey Methods and Theory, Volume II: Theory, John Wiley & Sons.

Knaub, J.R., Jr. (1992), "More Model Sampling and Analyses Applied to Electric Power Data," Proceedings of the Section on Survey 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. (1994), "Linear Modeling for Imputation in Establishment Surveys," Washington Statistical Society Seminar, Economics Section, June 9, 1994. (unpublished, but papers available from presenter/author)

Rao, Poduri, S.R.S. (1992), unpublished letters, Aug. - Oct. 1992, on covariances associated with three Royall and Cumberland model sampling variance estimators.

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. (1978), "Variance Estimation in Finite Population Sampling," Journal of the American Statistical Association, 73, pp. 351-358.

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.