A GENERALIZATION OF THE RATIO ESTIMATE WITH APPLICATIONS

Darrel W. Parke and Stephen B. Taubman, Federal Reserve Board

Occasionally in sample surveys when estimating a population total (Y) of a variable y_i, we also have available an auxiliary variable x_i, or covariate, which is known for each member of the population. A common estimator in such situations is

$$(1) \qquad \hat{Y} = (y'/x')X$$

where $y' = \sum_{i\in s} y_i$ and $x' = \sum_{i\in s} x_i$.

Let $B=Y/X$. In classical sampling, the properties of y'/x' as an estimator of B are investigated. The distribution of y'/x' arises from consideration of all possible sample outcomes given the survey design. To our knowledge, there do not exist general conditions under which y'/x' satisfies any sort of optimality conditions. In the context of sampling from an infinite population, however, it is shown in Cochran (1963) that $(\hat{y}/\hat{x})$ is the minimum variance unbiased estimator of β if y_i is generated by the following model

$$(2) \qquad y_i = \beta x_i + \varepsilon_i$$

$$E(\varepsilon_i) = 0 \quad Var(\varepsilon_i) = \sigma^2 x_i$$

$$E(\varepsilon_i \varepsilon_j) = 0, \; i \neq j$$

With this model, the generalized least squares estimator of β is the ratio estimator

$$(3) \qquad b_1 = y'/x'$$

which is also the maximum likelihood estimator when the ε_i are normally distributed.

We wish to consider a generalization of the above model in which we only require that the variance of ε_i is proportional to a power of x_i. Thus the model is

$$(4) \qquad y_i = \beta x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2 x^{2k})$$

where k can be any real number, although it has been nonnegative in all applications we have encountered. The likelihood function is

$$(5) \qquad L(y_1,\ldots,y_n;\beta,\sigma^2,k) = \prod_{i=1}^{n} (\sigma x_i^{k})^{-1}(2\pi)^{-n/2} \exp - \sum_{i=1}^{n} (y_i - \beta x_i)^2/2\sigma^2 x_i^{2k}$$

The maximum of the likelihood function is attained when

$$(6) \qquad \hat{\beta} = \sum_{i=1}^{n} y_i x_i^{1-2\hat{k}} \Big/ \sum_{i=1}^{n} x_i^{2-2\hat{k}}$$

$$(7) \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}x_i)^2 / x_i^{2\hat{k}}$$

and $\hat{k}$ is the solution to the equation

$$(8) \qquad \hat{\sigma}^{-2} \sum_{i=1}^{n} (y_i - \hat{\beta}x_i)^2 x_i^{-2k} \ln x_i = \sum_{i=1}^{n} \ln x_i$$

Special values of $\hat{k}$ yield familiar formulas for β. When $\hat{k} = 0$

$$(9) \qquad \hat{\beta} = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2$$

the ordinary least squares (OLS) estimator. When $\hat{k} = .5$,

$$(10) \qquad \hat{\beta} = \sum_{i=1}^{n} y_i \Big/ \sum_{i=1}^{n} x_i = y'/x'$$

the ratio estimator mentioned earlier. When k=1

$$(11) \qquad \hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} y_i/x_i$$

the mean of the ratios.

The asymptotic variances of the parameter estimates are

$$(12) \qquad Var(\hat{\beta}) = \sigma^2 / \sum_{i=1}^{n} x_i^{2-2k}$$

$$(13) \qquad Var(\hat{\sigma}^2) = 2\sigma^4 \sum_{i=1}^{n} (\ln x_i)^2 / nS(\ln x)$$

$$(14) \qquad Var(\hat{k}) = 1/2S(\ln x)$$

where $S(\ln x) = \sum_{i=1}^{n} (\ln x_i)^2 - (\sum_{i=1}^{n} \ln x_i)^2/n$

An Application

A sample of member banks will report to the Federal Reserve System the amount of their outstanding business loans at the end of each week. All member banks will report their total loans weekly and have reported their total and business loans once each quarter on the "call report". To investigate the estimation of a weekly series of business loans we take as our data base two consecutive call reports so that for each bank we have the values of

$$(15) \qquad \begin{aligned} y_i &= \text{business loans at bank i on second call date} \\ x_{1i} &= \text{business loans at bank i on first call date (PBL)} \\ x_{2i} &= \text{total loans at bank i on second call date (CTL).} \end{aligned}$$

(The second call date is a proxy for the weekly reports). The figures shown in Table 1 were compiled to aid us in choosing whether to use x_1 or x_2 as the covariate in the estimation of Y

and to select the form of the estimator (i.e. select k). The entries in the table are based on data from all member banks which are regarded as a sample from an infinite population. Some scaling would have to be done in order to calculate corresponding statistics for the actual weekly sample.

The estimated value of k is near unity when current total loans is the covariate. One interpretation of this finding is that the variance across banks of the fraction of business loans to total loans is nearly constant, that is, nearly independent of size of bank. A k of 0.644 for PBL indicates that, while the variance of business loans around the regression line increases with size of bank, the variance of the percentage changes in business loans from the preceding period decreases with size of bank. (The standard deviations of k are quite small—about 6000 observations went into the calculations).

The standard deviation of $\hat{\beta}$ is lower for current total loans than for previous business loans but, recalling that our ultimate objective is an estimate of the form $\hat{Y} = \hat{\beta} X$, the next line in the table is more relevant for choosing between CTL and PBL. These entries show that for estimating the mean aggregate business loans at a group of banks whose aggregate previous business loans and current total loans are equal respectively to those of the banks we have considered, the standard deviation of the estimate (382) using PBL as a covariate is only about half that obtained when using CTL as the covariate.

The next portion of the table examines the effects of using an inappropriate power of x in the estimation procedure. That is, given that $\hat{k}$ (1.013 for CTL and 0.644 for PBL) is the "correct" value of k, these entries are the standard deviations of

$$(16) \qquad \hat{\beta}(\ell) = \Sigma\, x_i^{1-2\ell}\, y_i / \Sigma\, x_i^{2-2\ell}$$

(This procedure is at least partly justified by the low standard deviations of $\hat{k}$ for the two covariates.) For PBL, use of the ratio estimator ($\ell = .5$) yields a standard deviation of $\hat{\beta}$ of .0032, about one quarter higher than that of the optimum estimator. The standard deviations of the ordinary least squares ($\ell = 0$) and means ratio ($\ell = 1.0$) are slightly more than twice as high.

The CTL data show a much different pattern. The standard devaition of $\hat{\beta}(1)$ is indistinguishable from the optimum ($\hat{\beta}$ (1.013)). But the standard deviations of the ratio and OLS estimates are eight and thirty times that of the optimum or mean ratio estimate. Thus, for the data considered here, when the ratio estimator is most appropriate, a moderate degree of precision is lost by using the OLS or mean ratio estimators, but when the mean ratio is appropriate, considerable loss is entailed when one of the other estimators is used.

It is well known (Scheffe (1959, Ch.10)) that estimation of variances in the general linear model is sensitive to violation of the homoscedasticity assumption. To illustrate the degree of this problem in our application, the last section of the table gives the estimated standard deviations of $\hat{\beta}(\ell)$. Here we are supposing that ℓ is used instead of k in the calculation of $\hat{\beta}$ and of its standard deviation. For $\ell = 0.0$, 0.5 and 1.5 in the CTL case these estimates are very low: .0015, .0018 and .0025 rather than the actual (i.e. given that $\hat{k} = k$) values of .0515, .0138 and .0102. In the PBL case the estimated standard deviations of $\hat{\beta}(.5)$ and $\hat{\beta}$ (1.5) are slightly more than one-half the actual values while the standard deviation of $\hat{\beta}(1)$ is over estimated by about one-third.

Finally, we wish to emphasize once again that the discussion has centered around the estimation of β, the coefficient of the regression equation. In many applications, the "parameter" of interest is $B = Y/X$. See Hartley and Sielken (1975), for example, for an explanation of the difference between the two approaches. Regarding $\hat{\beta}$ as an estimator of B, however, would not alter our qualitative results, although certain formulas would be changed. For example, when the sampling fraction is not negligible, we would modify the estimator by using

$$(17) \qquad \tilde{Y} = y' + \hat{\beta}(X - x')$$

rather than

$$(18) \qquad Y = \hat{\beta} X.$$

Conclusions

We have appealed to some infinite population concepts in order to attack a finite population sampling problem. We have found, somewhat counter to our intuition, that the optimum form of the estimator depends upon the covariate to be used, so that one may not arrive at the best available estimate of a population total if he treats the covariate selection and estimator selection problems separately.

We have also examined a bivariate ratio estimator (see Cochran (1963)) of business loans using PBL and CTL and found that it was only marginally better than a simple ratio estimator based on PBL. We plan to investigate whether and how the two estimators developed here can be combined.

Table 1

Statistics Relevant to the Selection of the Estimator for Business Loans

	Covariate	
	Current Total Loans (CTL)	Previous Business Loans (PBL)
$\hat{k}$	1.013	.644
$\hat{\sigma k}$	.0071	.0055
$\hat{\beta}$	.210	1.065
$\sigma_{\hat{\beta}}$	.0017	.0026
$X\sigma_{\hat{\beta}}$	784	382
$\sigma_{\hat{\beta}}$ using		
$\ell = 0.0$	.0515	.0066
0.5	.0138	.0032
1.0	.0017	.0059
1.5	.0102	.1377
2.0	.0878	.8964
Estimated $\sigma_{\hat{\beta}}$ using		
$\ell = 0.0$	.0015	.0016
0.5	.0018	.0017
1.0	.0017	.0099
1.5	.0025	.0760
2.0	.0019	.1820

[1] Cochran, W.G. (1963) Sampling Techniques. John Wiley and Sons, Inc., New York

[2] Hartley, H.O. and R.L. Sielken, Jr. (1975) A "Superpopulation Viewpoint" for Finite Population Sampling. Biometrics 31, pp. 411-422

[3] Scheffe, H. (1959) The Analysis of Variance. John Wiley and Sons, Inc., New York