# Discussion

Richard Valliant, Bureau of Labor Statistics, Washington DC 20212

Dr. Nascimento Silva is to be congratulated on organizing a session with a good set of related papers. The authors' work raises some timely questions given the recent surge in interest in model-assisted estimation. Rao and Singh give some unified theory for ridge regression and calibration methods plus provide some clear descriptions of computational algorithms. How to select variables for a regression estimator is a key issue addressed by Bankier, Houle, and Luc and by Nascimento Silva and Skinner. The comments here on the last two authors' work relate mainly to Nascimento Silva and Skinner (1997), which was the paper available when this discussion was given.

Much of the discussion of general regression, raking, and related methods emphasizes their ability to force sample estimates to equal benchmark values for different variables. Hitting census totals has some cosmetic appeal, but this fixation on matching benchmark totals seems misplaced. If our highest goal in a survey is to estimate well what we already know, then, perhaps, we are all being overpaid.

Rao and Singh give several ways of relaxing the benchmark constraints while also restricting the range of weights. Their results are very useful because they have devised a way of adaptively changing the constraints and maintaining asymptotic design consistency. They also provide detailed algorithms that lend themselves to programming in matrix languages like S-Plus™ or SAS/IML™ (also see Singh and Mohl 1996).

Adaptive procedures are sometimes difficult to develop theory to justify—take the stepwise regression procedures studied by Nascimento Silva and Skinner, for example. Rao and Singh, however, show how to do ridge regression with an adaptive choice of the ridge parameters or equivalently the tolerances on how closely the benchmarks are hit.

In their numerical study Rao/Singh observed some losses of precision using ridge-calibration methods compared to the regression estimator with no range restrictions on the weights. This is not always true, however. Jayasuriya and Valliant (1996) report a numerical study using household expenditure data in which restricted regression estimation yields coefficients of variation that are always less than or equal to those from unrestricted regression. We also did not observe any convergence problems using $L$=0.5 and $U$=2 in restricted regression with a substantially larger sample size than Rao/Singh had.

The model used by Rao/Singh is probably fairly weak, including only four demographic auxiliaries for each household. Although the types of predictors available for household populations tend to be limited, it would be nice to see how the alternative methods perform in a case where a richer set of auxiliaries can be used—in a business population, for example.

The Bankier, et.al., paper has goals of making sample estimates of person counts that match census population totals, and, at the same time, producing a set of household weights. The paper highlights one of the dilemmas faced when taking a modern-day census. If a sample is selected at the same time a census is done, there may be some pressure to force agreement between the census and sample estimates on characteristics collected by both. Achieving this kind of benchmarking for very detailed domains may dilute the gains that regression estimation can produce for more aggregated statistics.

The condition number (CN) reduction method used by Bankier, et.al., do not consider any response variable when deciding which constraints to retain. This contrasts with the variable selection methods studied by Nascimento Silva and Skinner that must have a $Y$.

CN is affected by the units of the variables, and, if the variables are on different scales, some of the best predictors may be lost as in the Nascimento Silva/Skinner example. In surveys where there are a few key response variables, the CN approach is not recommended because it does ignore the $y$'s. The Bankier, et.al., paper does highlight the need for careful numerical analysis when calibrating. Though collinearity among the $x$'s be unimportant when making predictions, as long as extrapolation does not occur (Gunst and Mason 1980, p. 310), most practitioners prefer models that are reasonably stable over time. Using nearly linearly dependent $x$'s seems risky.

The situations in the Nascimento Silva/Skinner paper and in Nascimento Silva and Skinner (1997) raise the question: What do we condition on when making an inference? The set of variables used in a regression model is a random event since different samples may lead to different sets. When should this source of randomness be accounted for in estimating a variance or constructing a confidence interval?

This situation bears some similarity to the measuring device example given by Cox and Hinkley (1974, p. 38). In that illustration, two instruments are available for a scientific experiment with much different characteristics. One is selected at random and the experiment conducted. It is known which instrument is used. An inference can be drawn either accounting for the initial random selection or not, but ignoring that step is justified since the random selection is ancillary. What to do about the selection of regression variables is less clearcut because the $Y$ and $x$'s used for subset selection may be earlier versions of the ones to be collected in a subsequent survey.

If the subset of auxiliary variables used for a regression estimator is selected based on a training dataset and then used until some periodic updating occurs, conditioning on that set of $x$'s seems like the sensible thing to do.

For example, suppose that a household survey is revised every 10 years after decennial census data are available for redesigning the sample. Assume that, as part of that revision, variable selection is redone but that the selected variables will be used for the next 10 years. Although the particular set of $x$'s certainly affects the bias and variance of the regression estimator, the fact that we might have picked a better or worse set of $x$'s seems irrelevant to making an inference using the particular set we did select. This argument seems more compelling as the survey period moves further away from the model development period. The possibility that we might, by chance, have chosen a different set of auxiliaries eight years ago, say, does not seem pertinent to constructing a confidence interval today.

A trivial example may better illustrate the conundrum. Suppose we have a population with the $y$-$x$ relationship pictured below. A simple random sample is selected and turns out to consist of the units in the ellipse. The usual variable selection methods probably will not pick up the dependence of $y$ on $x$, so we decide that the best estimator of the population mean is $\bar{y}_s$. Now, if we draw other samples and go through the variable selection procedures, quite often we will realize $x$ is a good predictor, and chose some sort of regression estimator. Of what relevance is this when we have decided to use $\bar{y}_s$ as the estimator in the one sample we have in our hands?

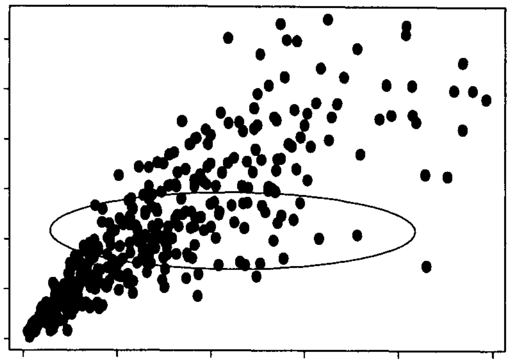

Figure 1 of Nascimento Silva and Skinner (1997) illustrates that there is a dependence of the *mse* on the number of auxiliary variables selected in their household population. The simulation study of Nascimento Silva and Skinner, on the other hand, does account for the random event of variable selection by re-selecting the $x$'s for each sample. When evaluating different methods like forward selection or best subset selection prior to settling on an estimator, accounting for that source of randomness seems perfectly reasonable. After a particular sample and set of $x$'s are selected, however, conditioning on that set is logical.

The issue of how much to account for in variance estimation is especially germane when using a replication variance estimator. It is a simple matter to repeat the variable selection process for each replicate sample, thereby reflecting that estimation step in an estimated variance. But, the wisdom of this seems doubtful. The set of selected auxiliaries can vary among the replicates. Thus, we could be computing a variance among point estimates that can be different from the one used in the full sample. The same questions also apply to adaptive procedures that can vary from one replicate to another. Suppose, for example, that one uses a post-stratified estimator but applies a collapsing rule if the sample size in any post-stratum dips below 20. Then applying the same rule to replicates might lead to different sets of post-strata in the replicates than in the full sample.

Confidence interval coverages in Tables 1 and 2 of Nascimento Silva and Skinner (1997) seem unusually poor, ranging from about 81% to 83% in Table 1 and from 81% to 89% in Table 2 for the subset selection strategies. Even the sample mean has less than nominal coverage at 91.8%. In Table 1 coverage percentages are poor even though the *mse* estimators are approximately unbiased. Perhaps this has to do with the fact that the test population contains units that may be correlated (in a model-based sense) since 426 households are used from only 20 enumeration areas. Skewed incomes may also be inducing a correlation between numerator and denominator of the $t$-statistics that would lead to poor coverage. Further investigation to discover the cause seems warranted.

### Author's Note

Any opinions expressed are those of the author and should not be construed as policy of the Bureau of Labor Statistics.

### References

Cox, D.R. and Hinkley, D.V. (1974), *Theoretical Statistics*, London: Chapman and Hall.

Gunst, R.F. and Mason, R.L. (1980), *Regression Analysis and Its Application*, New York: Marcel Dekker.

Jayasuriya, B. and Valliant, R. (1996), "An Application of Restricted Regression Estimation in a Household Survey, *Survey Methodology*, **22**, 127-137.

Nascimento Silva, P.L.D. and Skinner, C.J. (1997), "Variable Selection for Regression Estimation in Finite Populations, *Survey Methodology*, **23**, 23-32.

Singh, A.C. and Mohl, C.A. (1996), "Understanding Calibration Estimators in Survey Sampling," *Survey Methodology*, **22**, 107-115.