Confidence Interval Coverage in Model-Based Estimation

Wendy Rotz, Jinhee Yang, and Archana Joshee
Ernst & Young LLP, 1225 Connecticut Ave., NW, Washington, DC 20036

Key Words: confidence interval, model-based estimation, deep stratification

1. Introduction

In many business settings, the cost of reviewing a sample is extremely high, giving rise to the need for accurate estimates with narrow confidence intervals (CIs) using small samples. When there is a strongly related auxiliary variable, model-based estimation with deep stratification is a potential solution.

Generally, deep stratification reduces the sampling error and produces conservative confidence intervals. Through computer simulations, we explored whether there are situations when the confidence interval coverage is too conservative or not as robust as believed and found examples of both. We studied the effect of the population distribution, model fit, and degrees of freedom on confidence interval coverage.

2. Background

Corporations, such as chain stores and other businesses, may own or rent multiple buildings across the country. The fixed assets at each site (such as carpets, parking lots, light fixtures, and sign posts) are assigned to depreciation categories for tax purposes. For example, the pavement in a parking lot takes over 30 years to completely depreciate while carpet depreciates in just 5 years. There are several depreciation categories, such as 5 years, 7 years, 15 years and 39 years. Builders commonly assign all assets to long life categories. It is beneficial to the company, where appropriate, to reassign fixed assets to shorter depreciation categories.

Classification of assets into these categories is a costly process requiring engineers or architects to review blue- prints and visit the properties as well as lawyers or accountants to interpret tax laws and opinions. Ernst & Young, LLP took the lead among the large accounting firms to apply a statistical random sampling approach.

For cost-effectiveness, feasibility, and time constraints, samples need to be as small as possible. Yet for IRS acceptability and sound statistical practice, good precision and narrow confidence intervals are desired. Design-based sampling and estimation usually will not achieve enough precision in the sample size ranges that are feasible, but model-based sampling and estimation will.

The variable, total assets, usually proves to be a strong covariate and models based on assets generally have a small Mean Square Error (MSE) allowing much better precision than design-based approaches. We use the sampling approach of deep stratification to obtain representative samples and reduce variability.

3. Research Questions

With these very small samples, we questioned whether the model-based confidence intervals are properly formed. Specifically our questions were in three areas.

1) Our deep stratification methodology eliminates highly skewed samples. Therefore, we would expect our confidence intervals would be less likely to fail to contain the true values. However, could it be that the confidence intervals we create are too wide and much more conservative than necessary?

2) Small samples should and do have wider confidence intervals due to their smaller degrees of freedom and larger t-values. Are these confidence intervals wide enough – especially in the presence of greater variability? That is, will larger variability be properly reflected in a sufficient widening of the confidence interval or could there be instances of poor coverage in the presence of greater variability?

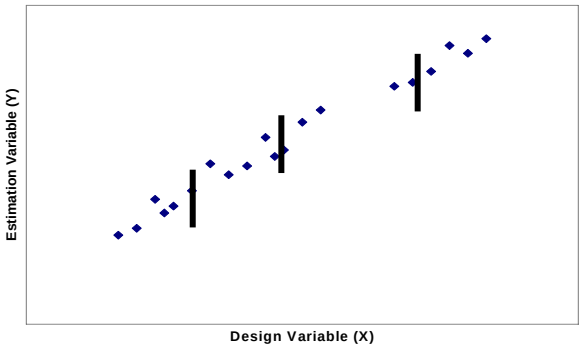
3) Typically after plotting the sample findings, we determine the type of model to build; making decisions regarding homo or hetero scedasticity, deciding whether to incorporate an intercept term, and assessing the need to compensate for curvature via a data transformation. Yet we make this determination on very few data points. What if we missed/ignored the underlying relationship and there are design flaws in the model? Would this be captured by the resulting confidence interval or could a modest design flaw cause poor coverage?

4. Deep Stratification

Deep-stratification is a random sample selection method that allows a small representative sample to be drawn.[1] Our design variable, X, is our covariate or independent variable, which typically is the total assets at a specific site. The estimation variable, or dependent variable, is the total assets belonging to a specific depreciation category.

The population is sorted by X and divided into equal sized strata according to counts. Equally sized random draws are made within each of these small strata. An illustration of the stratum cuts is shown in Figure 1. Note, of course, that the Y values are not actually known at the time of selection except in this simulation.

Figure 1. Deep Stratification Example



The resulting selections all have an equal probability of selection, meeting the definition of a simple random sample. However, unlike ordinary simple random samples, deep stratification samples cannot by chance have all high values or all low values in the sample. Therefore, deep stratification eliminates many of the possible simple random samples that are unrepresentative due to a disproportionate number of high and low values.

5. Research Method

We performed numerous computer simulations in SAS to explore confidence interval coverage in different settings. We generated several population data sets creating both an X and Y value for the entire population. Therefore, unlike reality, actual population Y values were known in these simulations and could be used in the measure of interval coverage.

We then drew multiple random samples from the same population using deep stratification. From each sample, we estimated Y, a total dollar figure, with model-based ratio techniques[2], created a confidence interval for the estimated value, and determined whether the actual value of Y was within the confidence interval. Confidence interval coverage was evaluated by the percent of confidence intervals containing the true value compared to the ascribed confidence level.

Because financial data is highly skewed, we used a gamma distribution to create X. Also, in our real world settings, there is a strong relationship between X and Y; usually R^2 is over 90%. Therefore, we created strong relations in our simulated population data, yet also considered some scenarios with larger variances.

With, the exception of when we were evaluating curvature or scedasticity assumptions, Y by design was created according to a linear relation to X with some random normally distributed heteroscedastic noise with a mean of zero and variance equal to a constant MSE times X. We tested a small, medium, and large value for the constant MSE.

In addition, to mimic our setting, Y was restricted to range between zero and X. By choosing appropriate slopes and using a portion of data that was distant from the origin, these restrictions were infrequently applied in most scenarios, so we initially believed our analysis would not be confounded by the occasional truncation of Y in most settings. We reconsidered the possibility of this confounding factor when we analyzed our results.

We created populations of size of 36, 180 and 720 records. We began with extremely tiny samples of just 9 and 18 records. Larger sample sizes may be tested in subsequent papers.

We drew ten thousand random samples for each scenario (where a scenario is a combination of population size, sample size, variance level, and relation of Y to X). For all ten thousand samples in each scenario we estimated Y using a heteroscedastic model:

$$Y = \beta X + \varepsilon$$

under the assumptions:

[1] Mary Batcher & Yan Liu (2003), Ratio Estimation of small samples using deep stratification, *Proceedings of the 2003 Joint Statistical Meetings*, Survey Methodology Section: American Statistical Association, Alexandria VA

[2] Lohr S. (1999) *Sampling: Design and Analysis*, Duxbury Press: Pacific Grove, CA, pages 81-83

i. $E(\varepsilon_i) = 0$ for all i.

ii. $V(\varepsilon_i) = X\ \sigma^2$ for all i.

iii. $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$

according to the standard formulas in Lohr .[3]

Finally, we calculated 80%, 90%, 95% and 99% confidence intervals and for each level we determined the actual coverage for the 10,000 samples drawn under each scenario.

6. Simulation Results

There is insufficient space available to present all of our findings, so we will share some of the more interesting and puzzling results in this paper.

Variance effect

First, we studied the effect of increasing variance. Figure 2 shows three sets of populations of size 180. The other size populations are similar, just differing numbers of records. The X values are the same in each scenario and the variance was set to be heteroscedastic.

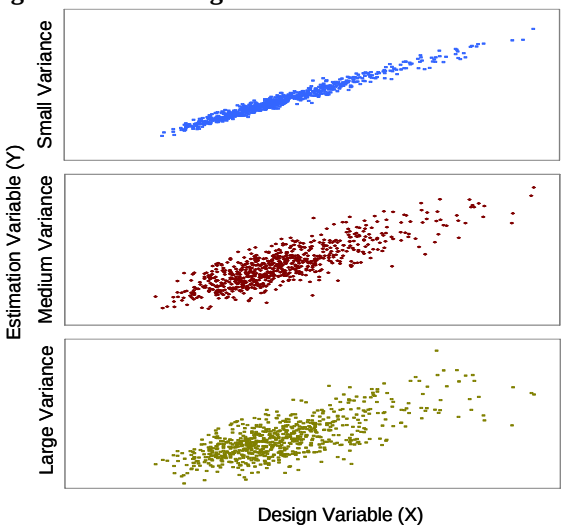
The relation between X and Y is the same with the exception that the MSE was increased in each scenario causing an increase in data dispersion. The three levels of MSE were 400, 4,900, and 16,000 respectively for the small, medium and large variance scenarios.

In practice we are most likely to encounter the small variance scenario. The medium variance is more than we normally see and the large variance scenario was studied to have an extreme case scenario.

These plots illustrate our first finding. We created our simulated population Y values with a heteroscedastic relation to X and therefore, expected to see the classic fan or funnel shape of heteroscedastic data. However, there is only slight evidence of this pattern in the last plot with the largest variance.

We determined this was due to the gamma distribution used to generate X. By design, as in our financial records, there are few instances of very large values of X, allowing fewer opportunities to observe the widely varying values in the tail of the distribution. When we applied the same formula for creating Y to a rectangular distribution of X, we observed the classic fan shape expected of heteroscedastic data, even for the smallest variance.

Figure 2. Increasing Variance Simulations



Each variance scenario was tested with two samples sizes of 9 and 18 records respectively out of populations of 36, 180, and 720 records. The results are presented in Figure 3.

Because the CI coverage demonstrated the same pattern for all confidence levels, we only present the 90% CI results in this paper.

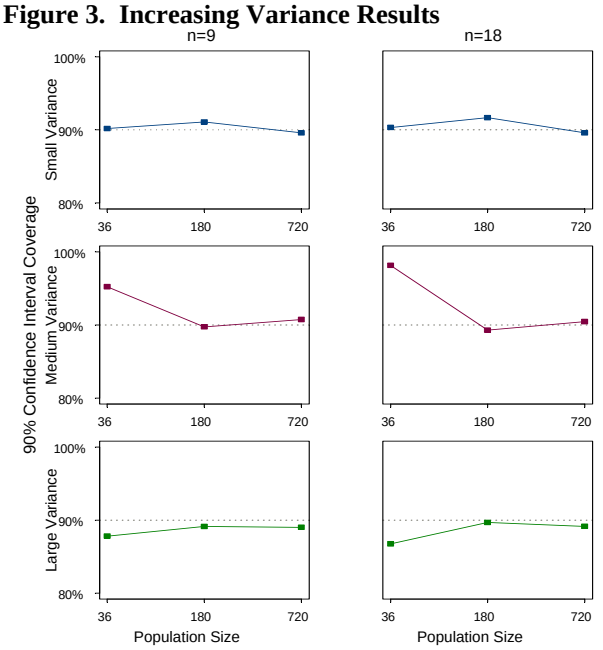
The y-axis shows the coverage with the reference line set at 90% where the coverage should be, as ascribed. The x-axis is the population size. The first plot in each row is for a sample size of 9. The second is a sample size of 18. The plots are ordered according to increasing variance.

The small variance scenario is what we expected. The coverage was very near or slightly above 90%. However, we did not attribute the slight increase to deep stratification because the pattern did not hold consistently.

In the medium variance scenario the confidence interval coverage is even higher, near 95%, for the small population. Better coverage was an unexpected result of increasing the variance. This small population does have the largest sampling fraction, but it did not consistently have better coverage than other populations.

[3] Lohr S. (1999) *Sampling: Design and Analysis*, Duxbury Press: Pacific Grove, CA, pages 81-83

Figure 3. Increasing Variance Results



In the large variance scenario, all coverage is below 90% in every scenario, with the smallest population showing the worst coverage. This puzzling finding caused us to revisit whether the truncation of Y could have caused these results. We did have more difficulty limiting the frequency of truncation, with the extremely large MSE.

Truncating Y could have caused less extreme values in the population than assumed by the model and therefore less extreme values available to sample resulting in narrower confidence intervals and hence poorer coverage. Although the instances of truncation were infrequent, we cannot dismiss their potential effect in light of these findings. This will be a subject of study in our next paper.

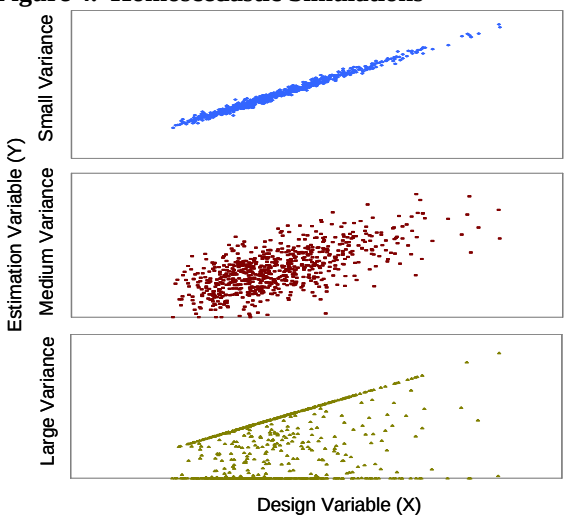
Still, we have demonstrated that while the larger variance did create wider confidence intervals, the resulting intervals were not wide enough. If this were a result of the truncations on Y, so that our model assumptions were not truly met, we need to consider that our real-life settings have the same practical limitations on Y and may not exactly fit the model assumptions either. Fortunately, however, we have much stronger relations between X and Y than was tested in the large variance scenario.
This was intended to be the series of scenarios where the model assumptions were met, yet we may have inadvertently introduced a truncation issue violating the assumptions. The next series of tests were on scenarios where the model assumptions were deliberately violated.

Scedasticity Assumption Effect

Next we tested the variance assumption effect if we estimated using a heteroscedastic model when the underlying data were actually homoscedastic.

Note that the homoscedastic data in Figure 4 looks quite similar to the heteroscedastic data in Figure 2 for the small and medium variances, leaving us to wonder whether we should ever trust a visual inspection to determine the variance assumption from a small sample. Note also that in this series, we allowed visibly severe truncation of Y in the largest variance setting.

Figure 4. Homoscedastic Simulations

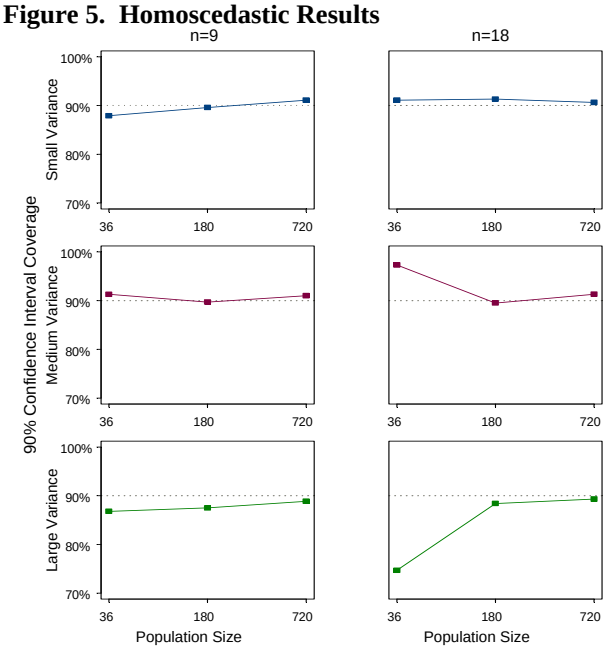


The resulting confidence interval coverage is similar to the findings above with a few exceptions. We found slight undercoverage with a sample of 9 from a population of 36 with the small variance and severe undercoverage (70%) from a sample of 18 in a population of 36 with the large variance. See Figure 5.

The sample of 18 records from a population of 36 is an interesting result. It is counter intuitive that increasing the sample size from 9 to 18, which is half of the population, would worsen the coverage. In addition, the increase in variance between the small and medium scenarios increased coverage, while a further increase caused the coverage to plummet.

One would expect that as the sample size increases, the confidence intervals would narrow, yet the coverage should stay the same. In addition, we would expect that
as variance increases, the confidence intervals would widen, while the coverage still remained the same.

Figure 5. Homoscedastic Results



Yet, in our simulations with the large variance, we found the confidence intervals narrowed too much when increasing the sample from 9 to 18. In addition, with the larger sample size of 18, the confidence intervals widened too much with the medium variance and not enough with the large variance.

However, it should be noted that now our truncation and scedasticty effects are clearly confounded here. We varied both from the previous set of scenarios. In addition, as seen in the large variance plot of Figure 4, the dispersion of Y about smaller values of X appears smaller than larger values due to the truncation.

Therefore, ironically the effect of truncation may have been to create heteroscedastic data after all. We need to perform more testing un-confounding with truncation, variance and scedasticity pattern unconfounded. Again, fortunately in modeling depreciation, we do not encounter such extreme distributions with the pattern of the large variance shown in Figure 4.

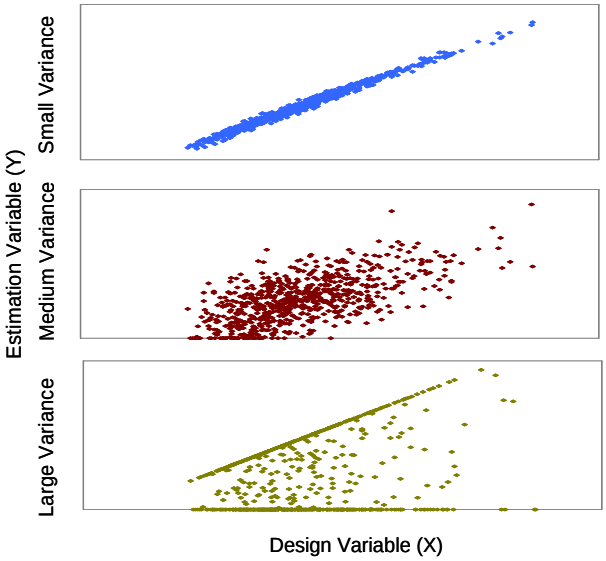
Intercept Effect

At first it may seem that it makes no business sense to include an intercept in our models because if there are zero assets, there cannot be a positive number assigned to a depreciation category. On the surface it may seem that a negative intercept could also produce nonsense because the regression line would cross the x-axis, placing a negative amount of Y into a depreciation category. This would result in total nonsense in our tax setting.

However, in statistical practice, the inclusion of an intercept will improve the model and an intercept often is included in the model regardless of its interpretation, especially when the data are far from the origin, as is often our case. Furthermore, the intercept can have a reasonable business interpretation as well. Often a property must be a sufficient size before finding any significant shorter term assets. Thus, rather than crossing into negative values of Y, the regression line could stop at the x-axis.

Unless guided by a pattern in the residuals, we typically do not include an intercept in our models. In this next set of scenarios we test the effect of failing to include an intercept term when the underlying data actually has one. The tested distribution is illustrated in Figure 6.

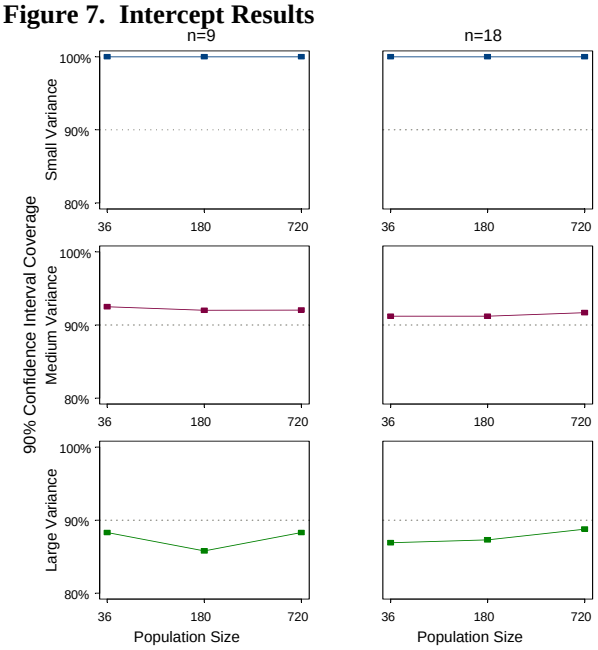
Figure 6. Intercept Simulation



We found that when we neglected the intercept, there were large residuals resulting in inappropriately wide confidence intervals. So wide for the small variance scenario, that coverage was near 100%. See Figure 7. Again coverage was reduced as the variance was increased, and again we are uncertain how much the truncation of Y may have confounded our results.

We concluded that even if we fail to recognize the need for an intercept in practice, our confidence intervals would probably be conservatively overstated because our real-life scenarios typically have smaller variances.

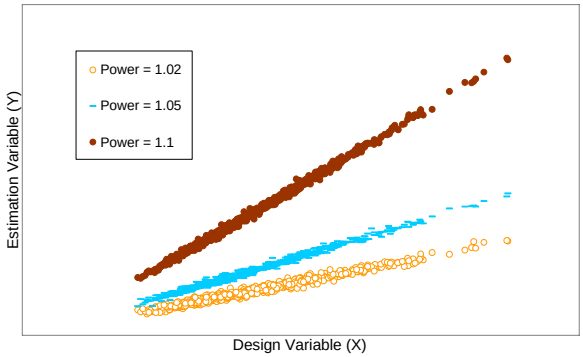
Figure 7. Intercept Results



Curvature

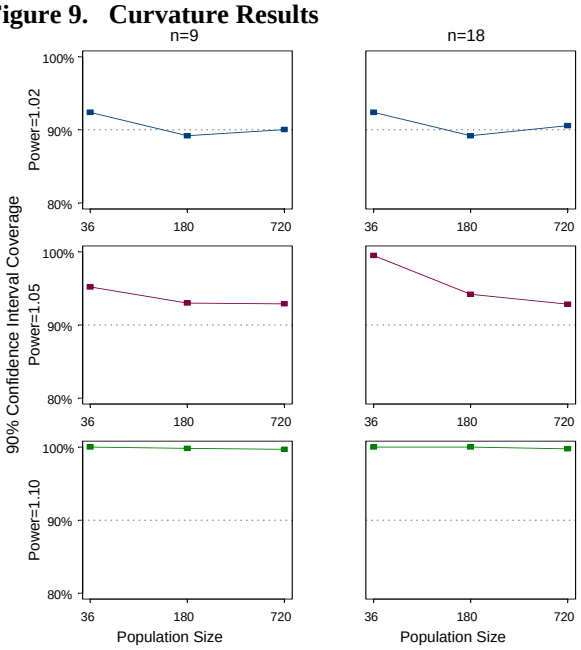
Finally we present the preliminary results of our curvature analyses. In this set of simulations, we explored the effect of neglecting to transform the data when there is some curvature present. That is, we created populations where Y is related to a power of X, $Y=BX^k+e$. See Figure 8.

Figure 8. Curvature Simulations



In our simulation, even a slight power caused large residuals resulting in wide confidence intervals which in turn caused over coverage.

Figure 9. Curvature Results



7. Conclusions and Next Steps

We have not yet addressed in our testing whether our deep stratification is causing confidence intervals to be too conservatively wide. To properly analyze this issue we need to un-confound the truncation of Y and the other issues tested as well as run similar scenarios both with simple random samples and deep stratification in order to ascertain whether any observed over coverage is attributable to the sample selection methodology. Due to time constraints, this analysis will be deferred to a subsequent paper.

We found instances of both over- and undercoverage when the relationship between X and Y fails to comply with our model and assumptions. The possible confounding of the truncation of Y was primarily an issue with the large variance scenarios. There was little if any truncation with the small variance scenarios and little truncation occurrence with the medium variance scenarios. Therefore, we comment only on these at this time.

For the small and medium variances we found that failing to include an intercept results in overly conservative confidence intervals in our simulations. More testing is needed to determine whether this pattern holds in general.

Also for the small and medium variances we found that failing to account for very modest curvature caused a startling amount of overcoverage. More testing with larger powers and fractional powers should be conducted.

Before concluding on the apparent reduction in coverage occasionally observed between the small and medium variance scenarios, we should assess whether a small truncation effect on the medium variance data could possibly be the cause of our findings. We also need to un-confound the issue of truncating Y before concluding on any of the substantial undercoverage examples we found with in our large variance scenarios.

In addition, until now, we have only studied two-sided coverage. We should also study one-sided confidence intervals to determine whether there is any asymmetry in coverage. We have not yet begun analyzing the distribution of the estimated values. This would be a high priority in our next paper especially in scenarios where we found over- or undercoverage of the confidence intervals.

Finally while we varied the simulated relations between X and Y, we assumed heteroscedasticity in all of our models for the purpose of estimation and construction of confidence intervals. Additional analyses with a homoscedastic assumption would be interesting.

Although there is still much work to be conducted, we have already found some surprising results. Our confidence intervals, while highly over covering, were at least conservative in the presence of modest amounts of curvature in the underlying data. Confidence interval coverage percentages from heteroscedastic models were nearly accurate or were conservatively over covering in the presence of low and moderate variance - whether the underlying data were actually heteroscedastic or homoscedastic.

The only egregious example of under coverage we found was with a small population and a large variance using a heteroscedastic model when the underlying data were actually the result of highly truncating homoscedastic data. If encountered, this kind of data would likely appear heteroscedastic in a sample plot. Statisticians should use caution when reporting confidence interval statements in settings where truncation may be present or is even exhibited in the sample findings.

In general, for the kinds of data we are most likely to encounter in the practice of sampling for tax depreciation, our confidence intervals can be robust under a wide variety of failures to meet the model assumptions.

References

[1] Alan H. Dorfman & Richard Valliant & Richard M. Royall (2000). *Finite Population Sampling and Inference: A Prediction Approach*, John Wiley & Sons: New York, NY

[2] Brewer K.R.W. (1999). Design-based or Prediction-based Inference? Stratified Random vs. Stratified Balanced Sampling. *International Statistical Review*, 67, 1, 35-47.

[3] Cochran, W.G. (1977), *Sampling Techniques*, 3rd ed., John Wiley & Sons: New York, NY

[4] Lohr S. (1999) *Sampling: Design and Analysis*, Duxbury Press: Pacific Grove, CA

[5] Mary Batcher & Yan Liu (2003), Ratio Estimation of small samples using deep stratification, *Proceedings of the 2003 Joint Statistical Meetings, Survey Methodology Section* ASA Alexandria, VA

[6] Royall R.M. & Cumberland W.G. (1981). An Empirical Study of Ratio Estimator and Estimators of Its Variance. *Journal of the American Statistical Association* 76, 373, 66-77.

[7] Royall R.M. & Cumberland W.G. (1981). The Finite-Population Linear Regression Estimator and Estimators of Its Variance – An Empirical Study. *Journal of the American Statistical Association* 76, 376, 924-930.

[8] Royall R.M. & Herson J. (1973). Robust Estimation in Finite Populations. *Journal of the American Statistical Association* 68, 344, 880-889.

[9] Sarndal, Swenson, and Wretman (1992) *Model Assisted Survey Sampling*, Springder-Verlag: New York, NY

[10]Tam S.M. & Chan N.N. (1984). Screening of Probability Samples. *International Statistical Review*, 52, 3, 301-308.

[11] Wendy Rotz & Archana Joshee, & Jinhee Yang (2006) Confidence Interval Coverage in Complex Model Based Estimation, *Proceedings of the 2006 Joint Statistical Meetings, Survey Methodology Section* ASA Alexandria, VA