COLLECTION FOLLOW-UP SCORE FUNCTION AND RESPONSE BIAS

Sanping Chen and Hansheng Xie[1]

ABSTRACT

Collection follow-up score functions based on weighted revenue are currently used in many business surveys at Statistics Canada. The objective is to maximize the coverage of the survey in terms of the primary variable of interest – total revenue, under the constraint of a limited follow-up budget. Such a strategy introduces a response bias toward large-contribution units in the sample, whose effect can be amplified through donor imputation during data editing.

This study examines the relationship among the three major financial variables, namely Operating Revenue, Operating Expenses, and Salaries, Wages and Benefits, as well as Sampling Revenue. It is noted that using weighted Sampling Revenue to sort units for follow-up is appropriate for the three major financial variables because of the latter's high correlations with Sampling Revenue. This study also obtains various estimates for the three financial variables, their bias and mean squared error (MSE) through simulating imputation and estimation processes. The results show a similar pattern for the three financial variables in terms of estimated bias and MSE, as well as the fact that Score Function can increase response bias toward large-contribution units and could have negative impacts on final estimates if the sampling distribution is highly skewed.

KEY WORDS: Follow up, Response bias, Score function.

RÉSUMÉ

Des fonctions de priorisation du suivi de la collecte basées sur le revenu pondéré sont actuellement utilisées dans plusieurs enquêtes auprès des entreprises. L'objectif est de maximiser la couverture de l'enquête en termes de la variable d'intérêt principale – revenu total, sous la contrainte d'un budget limité de suivi. Une telle stratégie introduit un biais de réponse vers les unités à grande contribution dans l'échantillon, dont l'effet est amplifié par l'imputation par donneur durant la vérification des données.

Cette étude examine les relations entre les trois principales variables financières, i.e. : le revenu d'opération; les dépenses d'opération et les salaires, traitements et bénéfices ainsi que le revenu à l'échantillonnage. Il est à noter que le fait d'utiliser le revenu d'échantillonnage pour trier les unités en vue du suivi est approprié par rapport aux trois variables financières principales puisque la corrélation entre ces dernières et le revenu à l'échantillonnage est élevée. Cette étude dérive également diverses estimations des trois principales variables financières, le biais et l'erreur quadratique moyenne (EQM) leur étant associé grâce à des simulations des processus d'imputation et d'estimation. Les résultats démontrent une tendance similaire pour les trois principales variables financières en termes de biais estimé et d'EQM ainsi que le fait que la fonction de priorisation de la collecte peut augmenter le biais de réponse vers les unités à grande contribution et qu'elle pourrait avoir un impact négatif sur les estimations finales si la distribution de l'échantillon est fortement asymétrique.

MOTS CLÉS : Biais de réponse; fonction de priorisation; suivi;

Sanping Chen (sanping.chen@statscan.ca) and Hansheng Xie (hansheng.xie@stats.ca), Statistics Canada, 17th Floor, R. H. Coats Building, Tunney's Pasture, Ottawa, Canada, K1A 0T6

1. INTRODUCTION

In business surveys at Statistics Canada there are significant costs to follow up non-responses and edit failures. It is important to obtain data from all sampled units but some units are more important than others due to their impact on the estimates. Therefore, Statistics Canada has developed follow-up score functions, usually based on the principal design variable which in most cases is the total business revenue. These score functions are currently used in many business surveys. The objective of the score function is to maximize the coverage of the survey in terms of the primary variable of interest – total revenue, under the constraint of a limited follow-up budget.

While such score functions help identify the important units and set priorities for follow-up, the strategy risks introducing a response bias toward high-revenue units in the sample. This effect is amplified through donor imputation during data editing. As more and more surveys start to utilize tax and other administrative data, the above response bias can be further exacerbated by response-based modelling and calibration of the administrative data. There is an acute need to examine this follow-up bias and to explore various ways of alleviating and correcting the effect of the bias resulting from the use of a collection follow-up score function.

In this paper, some preliminary results on the pattern of follow-up bias are presented based on several annual business surveys of the Services Industry conducted by Statistics Canada. These surveys all use Statistics Canada's general Business Register as the frame, the North American Industry Classification System (NAICS) as the industry classification, and the statistical establishment (a group of establishments for the most part) as the unit of interest.

Our emphasis is the impact of a follow-up scheme based on the primary design variable (total business revenue) on the bias of estimation of several major survey variables that may or may not be strongly correlated with the design variable.

2. SCORE FUNCTION

For each Services Industry survey covered by this study, a sample is selected at the establishment level using stratified simple random sampling without replacement and data are then collected either at the establishment level or at the company level, depending on the requirement for that survey. The process of data collection consists of designing questionnaires for each survey, mailing out questionnaires to respondents, and following up the collection units that are either non-responses or edit failures.

2.1 Method and Application

A score function is essentially a measure of importance of collection units based on their contribution to the coverage in terms of sampling revenue. The higher the score, the higher the priority for the unit to receive follow up. In each survey, the score function is applied to the group of collection units belonging to the same industry-geography (usually province/territory)-business size (revenue) stratum. Before collection, the score function assigns an initial *contribution score* to each collection unit within each group. This score is the collection unit's contribution to the *group coverage* and defined as

$$C_i = \frac{Sample\,Weight_i * Sample\,Design\,Revenue_i}{\sum_{Group} Sample\,Weight_i * Sample\,Design\,Revenue_i} \quad ,$$

where i refers to the i^{th} collection unit in the group.

The *contribution score* defined above is a weighted coverage rate of Sampling Revenue. With a target coverage rate set for each group, collection units within each group are sorted by this contribution score in descending order. (For

simplicity, the coverage rate is called *coverage* hereafter.) These ordered units are divided into two subsets using the target coverage as the cut-off. The units belonging to the top subset are assigned a priority score of one ("follow-up units"), with the remaining units in the bottom subset receiving a priority score of zero ("non-follow-up units").

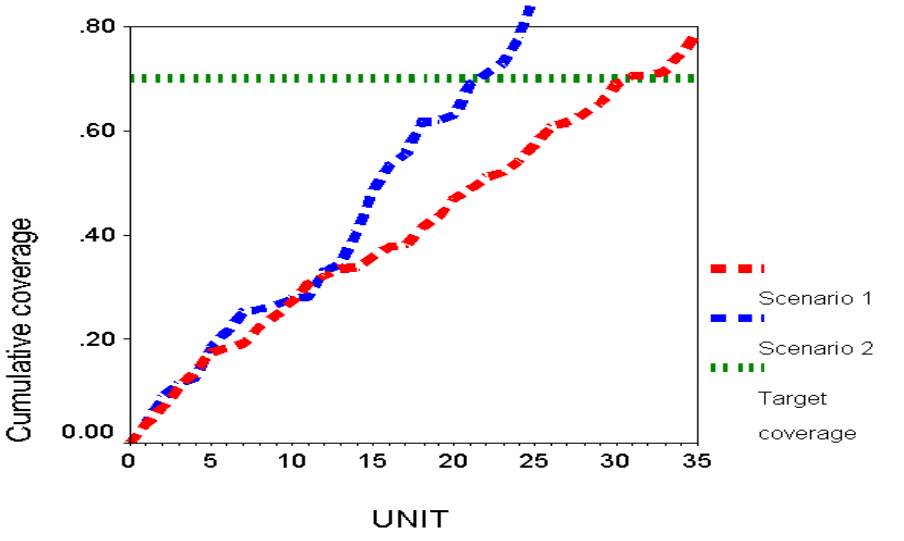
During the collection process, the units with a priority score of one are followed up. Then the *contribution score* is recalculated periodically based on the updated information. In the recalculation, all the units that have responded contribute to the coverage and are reassigned a priority score of zero; confirmed non-responses are also reassigned a priority score of zero since they will not provide any data; out-of-scope units are removed entirely from the process, and the priority scores of some units are promoted from zero to one to make up for the non-responses.

2.2 Statistical Meaning

The application of score function is a dynamic process that attempts to achieve a high coverage of the weighted revenue. It can be simulated as a stochastic process or a random walk. This process terminates when the target coverage is reached. Duration of this process is a random "stopping time".

The graph below shows two possible scenarios of the application. The two scenarios ended at different times, with different numbers of units collected when reaching the coverage of 70%.

Figure 1 - Random Walk and Stopping Time



3. RESPONSE BIAS STUDY

In the application of score function, the prominence of high-weighted response units could lead to a response bias. This bias could be further amplified through donor imputation during data editing. Without a study, it is unclear how significant the bias is and what relationship is between the bias and the follow-up strategy. Therefore, it is important to conduct a study to understand the impact of the score function on final estimates so that corrective measures can be taken to moderate the bias as required.

3.1 Data and Assumptions

A study was conducted using data from the 2002 Annual Survey of (Canadian) Arts, Entertainment and Recreation. The target population of this survey comprises companies that are primarily engaged in operating facilities or providing services to meet the cultural, entertainment and recreational interests of their patrons. To approximate the actual edit and imputation process, prior to the selection of a random sample, the population is classified into homogeneous groups by geography and sub-industry. Financial information is collected from the collection units at the company level in the sample and donor imputation is performed for missing information. Estimates of major financial variables are produced and published for sub-industries.
For this study, three major financial variables, namely, Operating Revenue, Operating Expenses, and Salaries, Wages and Benefits (SWB), are studied for three sub-industries using 1468 collection units. These collection units are distributed among the three sub-industries as follows: 289 units for Performing Arts (PA), 182 units for Spectator Sports (SS) and 997 units for Other Amusement and Recreation Industries (OAR).

The following assumptions apply to our study:

- The original data are used as a 100% response sample.
- Estimates produced from this original sample are used as pseudo true values for comparison purposes.
- For simulation samples, procedures of imputation and estimation are approximately the same as those used in production.
- All estimates are calculated using the estimation method for stratified simple random sampling without replacement.

3.2 Simulation Samples

The original collection units are first sorted by *contribution score* within each sub-industry. These ordered units are then divided into two subsets using a cut-off of 70% for coverage, namely the follow-up and non-follow-up subsets, respectively. A unit may not respond even though it is a large-contribution unit in the follow-up subset. Thus four follow-up options are set up for this study, which consist of different coverage for follow-up and non-follow-up subsets. Table 1 shows these four follow-up options.

Table 1 – Follow-Up Options

Option	Follow-Up Coverage	Non-Follow-Up Coverage
1	49%	21%
2	55%	15%
3	60%	10%
4	65%	5%

It can be seen that the follow-up on large-contribution units is enhanced gradually from Option1 (49%) to Option 4 (65%). In each of the four follow-up options, 500 simulation samples are randomly generated within each sub-industry. One should note that, with an overall coverage of 70% for each simulation sample, the number of units for each subset within each simulation sample is only determined by the coverage of the subset.

For each simulation sample, a randomly selected subgroup of units that covers 70% of the weighted revenue keeps the original variable values for estimation, while the non-selected units that have 30% coverage in total with their variable values imputed within each subset of a sub-industry using Statistics Canada's generalized imputation system.

3.3 Correlation Analysis

Since the weighted Sampling Revenue is used for sorting the collection units in production, it is necessary to look at the correlation among the three major financial variables and the Sampling Revenue. Table 2 contains the correlation matrix of the four variables.

Table 2 – Correlation Analysis on Financial Variables

	Operating Revenue	Operating Expenses	Salaries, Wages and Benefits	Sampling Revenue
Operating Revenue	1.00	0.98	0.80	0.74
Operating Expenses	0.98	1.00	0.85	0.71
Salaries, Wages and Benefits	0.80	0.85	1.00	0.62
Sampling Revenue	0.74	0.71	0.62	1.00

The above table shows that the Sampling Revenue has a relatively high correlation with the three major financial variables. This means that it is appropriate to use weighted Sampling Revenue for sorting collection units in production.

For studying the bias, a measure estimating the bias is defined as

Bias = Average of 500 Sample Estimates of a Variable – Pseudo True Value.

The correlation analysis is also conducted to study the correlation among the three bias estimates for the three variables. Table 3 contains the correlation matrix of the three bias estimates.

Table 3 – Correlation Analysis on Bias Estimates

	Bias Estimate for Operating Revenue	Bias Estimate for Operating Expenses	Bias Estimate for Salaries, Wages and Benefits
Bias Estimate for Operating Revenue	1.00	1.00	0.58
Bias Estimate for Operating Expenses	1.00	1.00	0.52
Bias Estimate for Salaries, Wages and Benefits	0.58	0.52	1.00

Tables 2 and 3 show that the three financial variables are strongly correlated with each other and so do the three bias estimates. This interesting relationship simplifies our study. As a result, one only needs to study one of the most important variables, Operating Revenue, from which the similar information may be obtained for the other two variables.

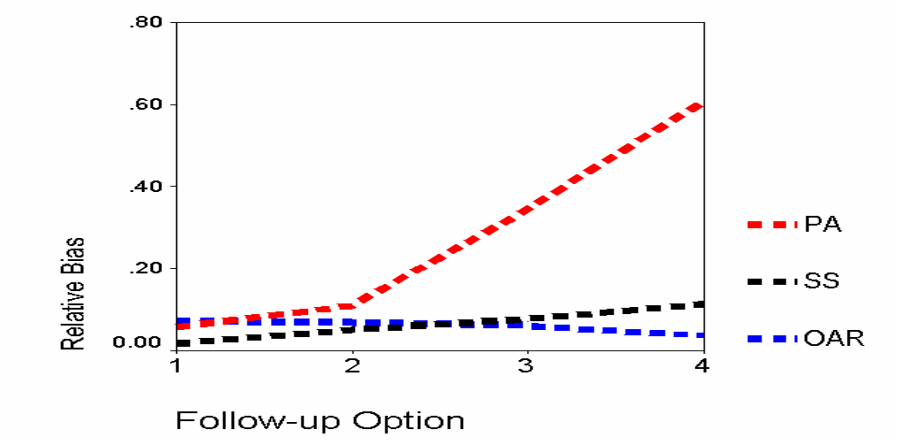
3.4 Results of Bias Study

For further examination of the bias, a measure estimating the relative bias is defined as

$$Relative\ Bias = |\ Bias|\ /\ Pseudo\ True\ Value.$$

In Figure 2, the Relative Bias for Operating Revenue is plotted versus the Follow-Up Option for the three sub-industries: Performing Arts (PA), Spectator Sports (SS) and Other Amusement and Recreation Industries (OAR).

Figure 2 - Relative Bias for Operating Revenue



It is clear from Figure 2 that, as the follow-up on large-contribution units is enhanced gradually, for Performing Arts and Spectator Sports the Relative Bias Estimates increase steadily, whereas for Other Amusement and Recreation industries the Relative Bias Estimates decreases slightly. One should note that, for Performing Arts, the Relative Bias Estimate goes up significantly from Option 2 to Option 3 (10.65% vs. 32.65%) and from Option 3 to Option 4 (32.65% vs. 58.32%).

The Mean Squared Error (MSE), defined as the sum of variance and squared bias, is a common measure of the total error of a survey statistic. In Figure 3, the square root of MSE for Operating Revenue is plotted versus the Follow-Up Option for the three sub-industries.

Figure 3 – Square Root of Mean Squared Error for Operating Revenue

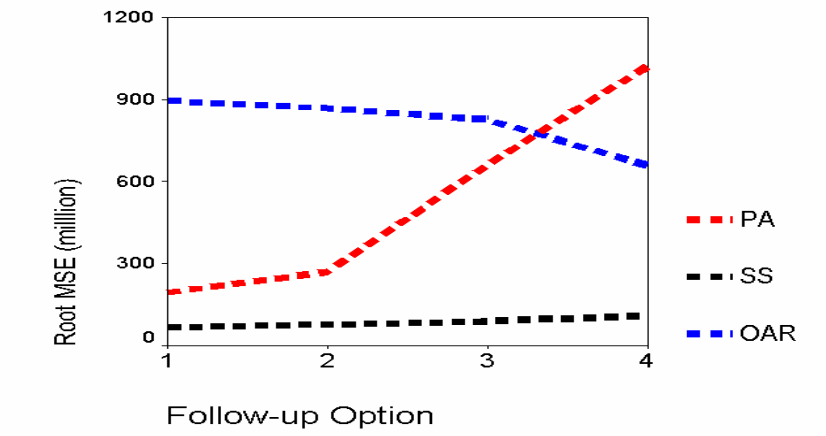


Figure 3 shows the patterns for the square root of MSE similar to those seen in Figure 2 for the relative bias. With the gradually enhanced follow-up on large-contribution units, the square root of MSE goes up for Performing Arts and Spectator Sports, and it goes down slightly for Other Amusement and Recreation industries. Especially for Performing Arts, the square root of MSE increases substantially from Option 2 to Option 3, and from Option 3 to Option 4.

For explaining why these patterns occur, two histograms are plotted in Figures 4 and 5. These two histograms display the distributions of the original collection units for the two respective sub-industries.

Figure 4 – Histogram of Operating Revenue (million) for Performing Arts

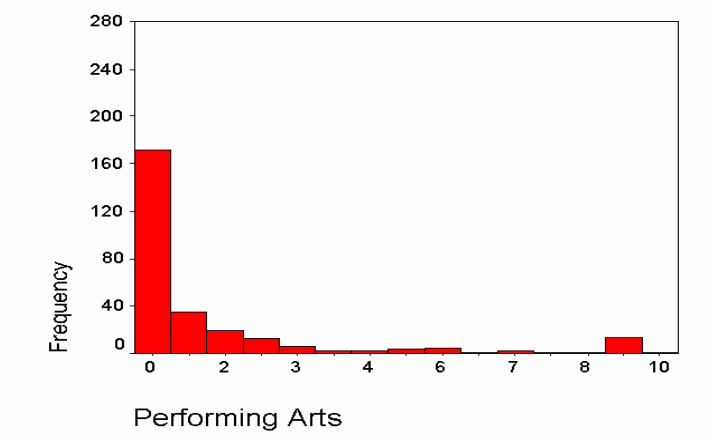
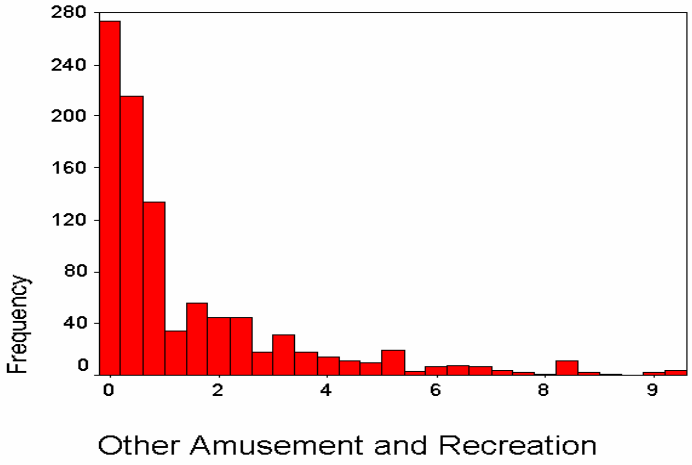


Figure 5 – Histogram of Operating Revenue (million) for Other Amusement and Recreation industries



These two histograms show that Performing Arts industry has a much more skewed distribution than Other Amusement and Recreation industries and its sample size is also much smaller. This means that, for Performing Arts industry, when the follow-up on large-contribution units is enhanced, there are very few donors available for imputation, which results in high bias and high MSE.

4. CONCLUSIONS

From the presented results, it is appropriate to use weighted Sampling Revenue for sorting collection units since this variable has a relatively high correlation with the three major financial variables. Because of their strong correlations, studying Operating Revenue can provide information about the other two variable estimates. However, the picture is unclear for variables weakly correlated with Sampling Revenue.

The study has shown that the score function can increase response bias toward large-contribution units if the sampling distribution is skewed. As a consequence, it can have negative impacts on final estimates when too few donors are available for imputation, due to, e.g., an extremely skewed distribution.

For alleviating the response bias, a corrective measure of providing scores at the business size stratum level has been taken in the application of score function. This can ensure that sufficient respondents are obtained from small and medium businesses. In a future study, the effects of this measure will be assessed. In addition, it is important to further explore other suitable corrective measures.

REFERENCES

Latouche, M. and Berthelot, J.-M. (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Journal of Official Statistics*, **Vol. 8, No. 3**, 389-400.

Philips, R. (2003). The Theory and Application of the Score Function for Determining the Priority of Follow Up in the Annual Survey of Manufactures. *The 2003 SSC Annual Meeting*, *Proceedings of the Survey Methods Section,* 121-126.

Pursey, S. (2003). Use of the Score Function to Optimize Data Collection Resources in the Unified Enterprise. *The 2003 SSC Annual Meeting*, *Proceedings of the Survey Methods Section,* 117-120.