SSC Annual Meeting, July 1995
Proceedings of the Survey Methods Section

ADJUSTING FOR NONRESPONSE IN THE REVERSE RECORD CHECK USING LOGISTIC REGRESSION AND CLUSTERING

S. Wu[1]

ABSTRACT

A method of using logistic regression and clustering to adjust for nonresponse in the Reverse Record Check is developed. This method provides an easy-to-implement approach to deal with nonresponse. In comparison to the current method for nonresponse adjustment, the new method has two major advantages: a) it performs better empirically in terms of bias and mean squared errors; b) it is fully automated, replacing several days of manual operation of collapsing initial weighting classes by several minutes of computation.

RÉSUMÉ

On met au point une méthode qui associe la régression logistique et la mise en grappes pour corriger la non réponse lors de la contre-vérification des dossiers. Cette nouvelle méthode permet de résoudre aisément la question de la non-réponse. Comparativement à la méthode utilisée à l'heure actuelle, elle présente deux avantages importants: a) elle donne de meilleurs résultats en ce qui concerne l'erreur systématique et l'erreur quadratique moyenne; b) étant complètement informatisée, elle permet de remplacer par quelques minutes de traitement plusieurs journées d'opérations manuelles consacrées au regroupement des classes de pondération initiales.

1. INTRODUCTION

The Reverse Record Check (RRC) is Statistics Canada's major study of undercoverage of persons and households in the Canadian Census. In the RRC, samples are drawn from frames that are independent from the current census and by matching the sampled persons with the census enumerations, the estimates of undercoverage are obtained. There are three main types of nonresponse in the RRC: non-identifiable, not-traced and non-classifiable. In this paper, we will concentrate on non-classifiables only. However, we believe that the proposed method can be modified to treat not-traced cases, as well.

Non-classifiable are the sample units for which we know that they were in scope for the census, but we were not able to clearly establish their enumeration status in the census. Namely, they are either enumerated or missed in the current census, but we do not know. This group represented 1% of the 1991 RRC sample.

The current method to adjust for non-classifiables is to re-distribute the (design) weights of nonresponses over the weights of responses who are traced and classified as "missed" or "enumerated". This is done within a large number of groups called "weighting classes". These weighting classes are formed by the combination of several variables which are thought to be closely related to the census result (missed or enumerated), including sample frame, sample province, the population of sampled area, age group, sex and other operational variables. Each weighting class is considered to be homogeneous in terms of chances of being missed.

There are three constraints on the weighting classes, namely:

1) There are separate weighting classes for each of the five replicates in the sample (used for variance estimation);
2) The number of observations in each class is no less than n; and
3) The proportion of nonresponses in each weighting class is no more than $p\%$.

Currently, n and p are 20 and 30 respectively. The initial weighting classes are collapsed together according to a set of rules until all the three constraints are met. For a more detailed description of the current nonresponse adjustment, see Burgess (1988), Garton (1982) and Royce (1991).

In general, the current method works well. However, there is room to improve. First, both the selection of variables to form the weighting class, and the rules to collapse initial weighting classes are subjective. Using our modelling approach, better variables can be selected and better collapsing rules can be established. Second, the selection of n and p is subjective. These too may be better selected. Third, the collapsing is done manually and is very time consuming. This in turn makes it very difficult to search for improvement. It is possible to

[1] Shiying Wu, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, 16-Q, R.H. Coats Bldg., Ottawa, Ontario, Canada, K1A 0T6.

automate the process to make it time efficient and to facilitate further improvement. Our proposed method will address these three problems.

The next section gives a brief description of the methodology used. The data set and the model fitting are described in Section 3. The evaluation of the results is reported in Section 4. Section 5 is a brief discussion on the results and some future research.

2. METHODOLOGY

The problem of non-response has been studied by many researchers and survey statisticians. Oh and Scheuren (1983) discussed the weighting class adjustment method. This is a post-stratified estimator using weighting classes as poststrata. This estimator is consistent if response propensities are homogeneous within weighting classes, where the response propensity is defined as the conditional probability of response given a set of predicting variables. Särndal and Swensson (1987) extended this approach to generalized regression estimators. One problem with this approach is that sometimes it is difficult to have the weighting classes homogeneous while maintaining a reasonable size. Raking ratio is another approach to adjust for nonresponse. Binder and Théberge (1988) showed that when the response propensities are multiplicative, then raking ratio estimators are unbiased. For a general discussion of design consistent estimation under nonresponse, see Singh, Wu, and Boyer (1995).

David *et al.* (1983) suggested the *response propensity stratification* approach. In this approach the response propensities are predicted by a model and then weighting classes are formed by grouping the predicted response propensities. Examples can be found in Judkins and Lo (1993). An alternative is discussed by Little (1986) and Särndal *et al.* (1992) where the weights of the respondents are inflated by the inverse of their estimated response propensities, without forming weighting classes. Such application can be found in Michaud and Hunter (1992). As Little (1986) pointed out, low values of estimated response propensity can inflate the variance of survey estimates excessively in this approach. Moreover it is not robust against the misspecification of the model.

Another approach is to model the characteristic of the interest, y, and form the weighting classes by grouping the estimated y given covariate $x, \hat{y}(x)$. Little (1986) called it *predicted mean stratification*. When comparing this approach with the response propensity approach, Little (1986) pointed out that predicted mean stratification has the virtue of controlling both the bias and the variance, while response propensity stratification only controls the large-sample bias. Eltinge and Yansaneh (1993) applied both approaches in their study and concluded that their data "provided relatively little power to distinguish between results of the two general cell-formation methods."

Our objective is to "properly" adjust for the non-classifiables so that the estimated number of missed persons should be close to those we would obtain if there were no non-classifiables observed. Obviously, it is important to reach this objective at the Canada level. Since the funds transferred from the federal government to the provinces are affected by the estimated numbers of missed persons in the provinces, it is also important to reach this objective at the provincial level.

The proposed adjustment for the non-classifiables is an application of the predicted mean stratification. Let x_i be a vector of variables of the ith sampled person, and p_i=Prob (the ith sampled person is missed $|x_i$). We first fit a logistic model

$$logit(p_i) = \log(p_i/(1-p_i)) = \alpha + x_i'\beta, \qquad (1)$$

where α is the intercept parameter, and β is the vector of slope parameters. The model was fitted with and without the weights. Both gave similar results. The result reported here is given by the model fitting without the weights. Stepwise selection of the covariates is used to obtain the final model.

From the final model, $\hat{p}_i$, the maximum likelihood estimate of p_i for each sample unit, is calculated. Then the $\hat{p}_i$'s are clustered into several clusters within each replicate-province combination. The clustering method used is called *nearest centroid sorting* (Anderberg, 1973). It selects initial cluster centers according to a prespecified (Euclidean) distance and each observation is assigned to the nearest center. Then it uses an iterative algorithm to update these centers and the clusters to minimize the sum of squared distances from the centers. Final clusters are obtained when the centers converge. The distances between initial cluster centers and the maximum number of clusters can be specified by the user. For a more detailed description of the method, see SAS/STAT User's Guide (Vol. 2, version 6, p. 824).

Within each cluster, the new weight of the ith sampled person is

$$w_i^* = w_i \frac{\sum w_i}{\sum_R w_i}, \qquad (2)$$

where $\sum$ denotes the sum over the cluster, and $\sum_R$ denotes the sum over the responses in the cluster.

3. DATA ANALYSIS

The Data. The RRC draws its samples from six different frames: census, birth, immigrant, missed, refugee

and permit holder, and health care file. Since the census frame accounts for more than 85% of total sample, we concentrate our discussion on this frame. Furthermore, because of the complication on design aspects of the two territories and some special features of undercoverage of Indian reserves, they are excluded from our study. With these exclusions, we have a sample size of 40,080. In addition, since only the sampled persons who completed the RRC questionnaire and were not found in Regional Office Operation, *Operation 3* (the first try to trace them), can be possibly classified as non-classifiable, any other sampled persons were excluded from the data. This gave us a total of 15,659 observations to fit our models.

Model fitting. Fourteen covariates were considered to fit the model, including age group, sex, marital status, census stratum, household size, mother tongue, tenure, type of dwelling, moving status and some other sampled person's characteristics. For a detailed list of these variables, see the Appendix. All the covariates used to fit the model are converted to groups of dichotomous variables except for census stratum and moving status.

The result of a preliminary model fitting indicated that there are six variables that are relatively important. They are:

MS: married or age < 20,
MALE: male,
AGE2: 20 ≤ age ≤ 24,
APT: the dwelling is an apartment,
ADDCDUH and ADD.

The last two variables characterize the Census Day address of the sampled person. See the Appendix for their definitions. Because of the importance of these six variables, all the second order interactions involving MS, MALE and AGE2, and the third order interaction AGE2*MALE*MS were considered in the model fitting.

The ten provinces were divided into 5 regions: East, Québec, Ontario, West and British Columbia. The provinces within each region were thought to have similar characteristics, in terms of undercoverage and its relation to the above covariates. Therefore, it was thought that fitting a model for each region should give us a reasonably robust result (as opposed to fitting a model to each province), and yet allowing differences between the regions.

To reduce the influence of sample variation on model selection, we decided to select a small set of common covariates that are important for the regions. To this end, stepwise regression procedure was used to fit Model (1) with the 6 relatively important variables and the interactions mentioned above, for each of the five regions. The significance levels for entry into the model, and for staying in the model were both set at 0.05. This choice was somewhat arbitrary. Yet it seemed satisfactory in the sense that it did give us neither too few nor too many covariates in the selected models (5 to 11 covariates were selected for the regions).

It turned out that MS, MALE and ADDCDUH are important since they were selected for at least four of the five regions with high significance levels. Therefore it was decided that these three covariates should be in the final models. Next, to see what other variables are important given these three covariates, the models were fitted again using stepwise regression with the three covariates forced into them. The result showed that covariates ADD, APT, P1 (see Appendix 1 for definitions) and MS*APT were more important than others, since for at least 2 regions, each of these covariates is significant at level of $p <$ 0.005. Hence, we decided that these four covariates should also be in the models. The cut-off level (seven covariates, rather than eight, for example, were selected) is somewhat subjective. However, experience tells us that too many covariates will make the models more sensitive to sample variation. An experiment was carried out for models with nine covariates. A cross validation study showed that it did not perform as well as the seven-covariate model we selected. In addition, for models with more than seven covariates, some of our replicates used in the cross validation experience singularity problems.

In summary, covariates MS, MALE, ADDCDUH, ADD, APT, P1 and MS*APT were selected for the final models, representing marital status, sex, Census Day address, living in an apartment, position in the census questionnaire, and the interaction between marital status and living in an apartment, respectively. The final model can be written as

$$logit(p_i) = \alpha + \beta_1 MS + \beta_2 MALE + \beta_3 ADDCDUH + \beta_4 ADD + \beta_5 APT + \beta_6 P1 + \beta_7 MS * APT. \quad (3)$$

A more objective approach to select the covariates is to use subset regression combined with cross validation to minimize a loss function such as the sum of squared errors in the estimated provincial totals. But that would require much more computational resources. For practical purposes, we feel that the proposed approach is satisfactory.

Using the seven selected covariates, the logistic regression model was fitted again for each of the five regions. The estimated p_i's were produced for the responses and non-classifiable sample units.

Clustering. Before clustering the estimated p_i's, the difference in the weights carried by the observations has to be considered. For example, a sample in the age group of 20 to 24 has a weight 50% less than those of others. If a non-classifiable with large weight is assigned to a small cluster where every response has a small weight, then the adjusted weights of the responses will increase drastically. This will result in a distortion in the distribution of the estimated undercoverage over age, and over other

variables. We would like that the weights in a cluster are as homogeneous as possible, provided that the clusters are not too small to make the estimates unstable. Apart from age groups, observations in different sampling strata also have significantly different weights, but to a lesser extent. Hence, in general, sample units in the age group 20-24 should not be clustered with sample units in other age groups, and sample units from different strata should not be clustered together. However, if the above rules are followed strictly, sometimes we end up with clusters with one or a few observations or clusters with no response, making the reweighting difficult. To obtain robust estimates, a cluster should have more responses than non-responses and its size should not be too small (say, > 15). The solution to this problem is to make a compromise: when we have no choice but to collapse a cluster with clusters of other strata or age groups, we collapse it with other clusters over strata first, and then over age groups if needed. This is to reflect the fact that the weights in different age groups vary more. The strategy was realized by adding 20 to the $\hat{p}_i$'s in the age group 20 to 24, and adding a different multiple of 2 to the $\hat{p}_i$'s in different strata. Note that the original $\hat{p}_i$'s were between 0 and 1. The addition would force the clustering procedure to separate the samples from different age groups and strata whenever possible.

After the addition, the nearest centroid sorting clustering procedure was applied to the $\hat{p}_i$'s within each replicate-province combination. The distance between initial clusters was set to be 0.06 and the minimum size of clusters was chosen as 16. These parameters were chosen by trial-and-error. We only tried a few different values on each parameter. More extensive experiments could be carried out when time permits. Nonetheless, these choices gave us reasonable results.

The number of clusters in each replicate-province combination was usually 3 to 7 but sometimes more and sometimes less. Within each cluster, the weights were revised by equation (2). Since the true undercoverage rate is unknown and hence cannot be compared with, the estimates obtained from the new weights are not reported here. The evaluation of this method is done by cross-validation, reported in the next section.

4. EVALUATION

In the RRC, a post-stratification adjustment is applied to the nonresponse-adjusted weights to obtain a set of final weights. The RRC estimates are then calculated by summing these final weights over the appropriate domains. The purpose of the post-stratification is to reduce the potential sampling bias.

In our cross validation, both the before post-stratification results and the after post-stratification results are compared. The before post-stratification results isolate the effects of different nonresponse adjustments, while the after post-stratification results show their impacts on the final estimates. However, since the two types of results are very similar, only the before post-stratification results are reported here.

First, we deleted all the non-classifiable sample units from the data set to obtain a set of "standard" weights (before post-stratification). From these standard weights, the estimated number of missed persons was produced for Canada and each province. These estimates were later used as a standard to compare the results of different methods. Note that these estimates are not comparable to the RRC estimated number of missed persons for Canada and the provinces.

Next, with the non-classifiables deleted, the data set was randomly divided into 20 (nearly) equal groups. A replicate data set was created by setting one of the groups to non-classifiable. This process was repeated ten times using different groups to obtain 10 different data sets. For each method, a set of new weights was calculated from each of the 10 replicate data sets. Therefore, we produced 10 sets of new weights for each of the current methods and the new method. Note that the fitted regression coefficients are different from one data set to the other. From each set of the new weights, the estimated number of missed persons was produced for Canada and each province. Since we had 10 replicate data sets, 10 sets of estimates were produced for each method. Let T be the estimated total using the standard weights, and let T_j be the estimated total using the weights calculated by a method with replicate j, $j=1, ..., 10$. Then the mean bias (over the 10 replicates) for that method is defined by $\sum_{j=1}^{10} (T_j - T)/10$ and the mean squared errors (MSE) is defined by $\sum_{j=1}^{10} (T_j - T)^2/10$. Table 1 shows the mean biases and the mean squared errors calculated for the two methods.

From Table 1, we observed the following.

At the Canada level, the new method performed better than the current method in terms of the mean bias and the MSE. Compared with the current method, the new method reduced the MSE by 23%. It also produced a much smaller absolute mean bias.

At the province level, the new method produced a smaller MSE than the old method in nine of the ten provinces. The new method also produced smaller absolute mean bias in seven of the ten provinces.

Another concern is what impact the new method might have on the estimated variances or the standard errors of the estimates in Table 1. The means and the standard errors of the standard errors (of the estimates in Table 1) were calculated and examined for the two methods. The result shows the standard errors (of the estimates in Table 1) produced by the two methods are

very similar and very stable. Since no significant difference between the two methods is observered, the actual numbers are not presented here.

Table 1: The Mean Bias and the MSE for the two Method before post-stratification.

MEAN BIAS (MSE)	CURRENT METHOD		MODEL (3)	
Newfoundland	-46	(12794)	-2	(8910)
P.E.I.	1	(427)	2	(337)
Nova Scotia	82	(55736)	81	(38979)
New Brunswick	-185	(62338)	-69	(25185)
Quebec	95	(999392)	-114	(1049301)
Ontario	-311	(1308535)	75	(679392)
Manitoba	124	(53046)	150	(47107)
Saskatchewan	-250	(79753)	-208	(58249)
Alberta	58	(119197)	5	(108186)
British Columbia	-241	(282025)	-86	(248742)
Canada	-674	(603350)	-167	(1235377)

5. DISCUSSION

We have proposed a new method to adjust for nonresponse. In the limited cross-validation study, the cross validation using the real data shows that the new method performed better than the current method in terms of bias and mean squared errors. In terms of variance, the new method performed about the same as the current method. In addition, the new method eliminated the time consuming manual operation of collapsing initial weighting classes, and automated the whole process. This replaced several days of manual operation by several minutes of computation. Such a feature itself is important since it will improve the timeliness of census products.

Eltinge and Yansaneh (1993) suggested to use the j/kth quantiles of the $\hat{p}_i$'s to form the weighting classes, $j=1, ..., k-1$. They also suggested $k=5$ may give most of the feasible bias reduction. However, for predicted mean stratification, Little (1986) pointed out the weighting classes should be chosen to maximize the ratio of between-to-within-class variance of y to reduce the variance of the estimates. Given the number of classes, this is equivalent to minimizing the within-class variance. This motivated us to use the clustering method that minimizes the within-cluster variance. The number of clusters is controlled by the distance between initial cluster centers. This is reasonable because it provides a control of the cluster width, and the bias reduction is achieved by the homogeneity of $\hat{p}_i$'s within the clusters. We think this provides a reasonable balance among bias reduction, variance reduction and robustness against estimation error of $\hat{p}_i$'s. Based on the above consideration, we expect that this approach would perform better than the quantiles approach. But, an empirical comparison is needed to support this speculation. Note that the cluster procedure can also be applied in the response propensity stratification approach.

In cross validation, we generated nonresponse randomly. In reality, however, it is likely that nonresponse is associated with some of the covariates. Hence other ways of generating nonresponse should also be considered.

In the future, the selection of the covariates in the model and the choice of the parameters in the clustering could be carried out more objectively by minimizing a loss function. Since the whole process is automated, a grid search for optimal parameters is possible. Also, the proposed method needs to be modified so that nonresponse in other frames, territories and Indian reserves could also be handled.

With appropriate modification of our models to handle multiple responses, the proposed method can be used to adjust for not-traced cases. This needs further research.

REFERENCES

Alexander, C.H. (1987). "A class of methods for using person controls in household weighting", *Survey Methodology*, 13, 183-198.

Anderberg, M.R. (1973). *Cluster Analysis for Applications*, New York: Academic Press.

Binder, D.A., and Théberge, A. (1988). "Estimating the variance of raking-ratio estimators", *Canadian Journal of Statistics*, 16, 47-55.

Burgess, R.D. (1988). "Evaluation of Reverse Record Check estimates of undercoverage in the Canadian Census of Population", *Survey Methodology*, 14, 137-156.

David, M., Little, R., Samuhel, M., and Triest, R. (1983). "Nonrandom nonresponse models based on the propensity to respond", *Proceedings of the Business Economics Statistics Section*, American Statistical Association, 168-173.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). "Generalized raking procedures in survey sampling", *Journal of the American Statistical Association*, 88, 1013-1020.

Eltinge, J.L., and Yansaneh, I.S. (1993). "Weighting adjustments for income nonresponse in the U.S. Consumer Expenditure Survey", Technical Report No. 202, Department of Statistics, Texas A&M University, College Station, Texas.

Garton, B. (1982). "1981 Reverse Record Check estimation methodology report", (Unpublished paper).

Judkins, D., and Lo, A. (1993). "Components of variance and nonresponse adjustment for Medicare Current Beneficiary Survey", *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Little, R.J.A. (1986). "Survey nonresponse adjustments for estimates of means", *International Statistical Review*, 54, 139-157.

Michaud, S., and Hunter, L. (1992). "Strategy for minimizing the impact of nonresponse for the Survey of Labour and Income Dynamics", *Proceedings of Symposium 92: Design and Analysis of Longitudinal Surveys*, Statistics Canada.

Oh, H.L., and Scheuren, F. (1983). "Weighting adjustments for unit nonresponse", in *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies* (W.G. Madow, I. Olkin and D. Rubin, eds.), 143-184. New York: Academic Press.

Royce, D. (1991). "Sample Design of the 1991 Reverse Record Check", (Unpublished), Statistics Canada.

Särndal, C-E., and Swensson, B. (1987). "A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse", *International Statistical Review*, 55, 279-294.

Särndal, C-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Singh, A.C., Wu, S., and Boyer, R. (1995). "Longitudinal survey nonresponse adjustment by weight calibration for estimation of gross flows", *Proceedings of the Survey Research Methods Section*, American Statistical Association.

APPENDIX: VARIABLES CONSIDERED IN THE MODEL FITTING

Marital status:

MS: married or age < 20.

Age group: age is divided into four groups: 0-19, 20-24, 25-34, and 35 and up.

AGE1: $0 \leq$ age ≤ 19;
AGE2: $20 \leq$ age ≤ 24;
AGE3: $25 \leq$ age ≤ 34.

Census frame stratum: defined by method of enumeration, population, urban/rural, and indian/non-indian.

STRATUM: 1 mail back, 10,000-29,999, urban, non-indian;
2 mail back, 30,000-99,999, urban, non-indian;
3 mail back, 100,000-499,999, urban, non-indian;
4 mail back, 500,000+, urban, non-indian;
0 others.

Sex:

MALE: male.

Census Day address of the sampled person:

ADDSPR: where the respondent was reached;
ADDCUH: the current usual home of the sampled person;
ADDCDUH: the usual home of the sampled person on Census Day;
ADD: known but none of the above.

Whom the sampled person was living with at the time of the RRC:

ALONE: alone;
WITHCF: census family;
WITHREL: relatives other than the census family.

Who is the sampled person:

P1: Person 1 in the Census;
P1CF: a member of the census family of Person 1 in the Census;

The household size of the sampled person:

HHLD1: single person household;
HHLD6: household size ≥ 6.

Whether the respondent is a member of the sampled person's census family:

CFRESP: yes.

Mother tongue:

MTO: mother tongue is neither English nor French.

Tenure:

OWNED: owned the house;
RENT: rented the house.

Dwelling type:

APT: the dwelling is an apartment;
MOBILE: the dwelling is mobile.

Moving status (0-2):

MOVED: known and has not moved; not known; known and has moved.

Linked status of RRC data to tax file:

LINK1: any member of the sampled person's family was matched;
LINK2: linkage was not attempted.