**Prediction Error Estimation of the Survey-Weighted Least Squares Model under Complex Sampling**

Retha Luus, University of the Western Cape, rluus@uwc.ac.za

Ariane Neethling, Stellenbosch University, ariane\_neethling@yahoo.com

Tertius de Wet, Stellenbosch University, tdewet@sun.ac.za

**Abstract**

*Many large-scale surveys make use of a complex sampling design where each observation unit is assigned a sampling weight which is developed over different stages. Survey-weighted least squares modelling (SWLS), the linear modelling of a continuous response based on its relationship with a number of covariates, correctly accounts for this complex sample design. One of the objectives of statistical modelling is the prediction of a future response. As such it is of importance to determine how well the selected model performs in the prediction of a future response. Cross-validation methods have long been used for this purpose under i.i.d data modelling, but not for the modelling of CS data. This paper introduces cross-validation for the evaluation of SWLS models based on the model’s prediction error. The investigation of the performance of the different prediction error estimation methods are evaluated through a simulation study. The Income and Expenditure Survey 2005/2006 of Statistics South Africa will form the basis of the analysis. The simulation study will also investigate whether the SWLS model’s predictive performance is improved through the truncation of outlier sampling weights. For this purpose two new thresholds, viz. the 1.5IQR and Hill, will be introduced.*

**Keywords:** Cross-validation, survey-weighted least squares, sampling weights, benchmarking, trimming

**1. Introduction**

Consider a finite population of size with an -vector of responses, , and predictor variables, , with , where represents the value of the th predictor variable for the th population element. Due to its ubiquity in applied statistics in the modelling of a response as a function of covariates, the model often used to define the relationship between the response and the predictors is assumed to be a linear model, i.e. , where is an matrix of population predictors, is a -dimensional vector of unknown regression coefficients and (Lohr, 2010).

One of the aims when statistically modelling data is to develop an accurate model that can be used to predict the response, based on this relationship, at a given covariate. The prediction error (PE) of a model is often used as a measure of its performance, i.e. how well, on average over a set of data, does the model predict the response. Naturally using the same sample data to develop and assess the model will give a distorted impression of the model’s predictive capability. However, future observations are unknown and as such not available for assessment of the linear model (James, et al., 2013). A well-known method to estimate the PE for independent and identically distributed (i.i.d.) data, is cross-validation. In this paper it will be adapted for complex sampling (CS) data and then used to estimate the PE of the survey-weighted least squares model.

Consider a sample selected through a stratified two-stage cluster design whereby a population has been stratified into strata and within stratum there are primary sampling units (PSUs) of which PSUs are selected. Let the ()-th selected PSU contain secondary sampling units (SSUs) and suppose a sample of SSUs is selected from this PSU, and . Each unit in this complex sample (CS) is assigned a design weight, ,, calculated as the inverse of the inclusion probability of the ()-th unit and a value indicating the number of population units represented by this sampled unit. The design weight is adjusted to compensate for any non-response in the sample and finally it is benchmarked, through the methods of calibration and integrated weighting, to the known population totals of certain auxiliary variables to ensure that the sample is well representative of the target population. After these weight development stages have been completed the sampling weights, ,, are obtained (Lohr, 2010; Neethling & Galpin, 2006).

When developing unbiased estimators of general unknown parameters from CS data, the variation in sample selection and inclusion probabilities necessitate the inclusion of these sampling weights (Heeringa, et al., 2010). Linear modelling of CS data, or survey-weighted least squares modelling (SWLS), does exactly this.

The weight development process described here could result in outlier sampling weights which, since the weights are included in the estimation of the SWLS model, could inflate the variability within the sampling weight distribution and hence have an adverse effect on the precision of the inference results. It has thus been proposed that the sampling weights be trimmed or smoothed to reduce this variability. Various procedures for doing this have been proposed in literature. Recent research by the authors have introduced two new weight trimming thresholds, namely the 1.5IQR and Hill thresholds, that performed very well in simulation studies (Luus, 2016). These will form part of the simulation study in section 3.

The purpose of this paper is to estimate the prediction error of the SWLS model using cross-validation. The next section briefly describes the different cross-validation approaches followed by their adjustment for use in the PE estimation of the SWLS model. Section 3 formulates the simulation study and introduces the data set to be used, i.e. the 2005/2006 Income and Expenditure Survey (IES) of Statistics South Africa. In section 4 the results obtained from this analysis are presented and discussed and, finally, conclusions and areas for further research are given in section 5.

**2. Prediction error estimation using cross-validation**

In CS the cross-validation will be carried out in each stratum since strata are considered to be independent non-overlapping subgroups into which the population has been divided for sampling purposes. Furthermore, the units within each stratum that are to be divided into training sets and a test set will be the PSUs, the first level sampling units within each stratum. The reason for this is to ensure that the structure within the PSUs remains preserved.

General -fold cross-validation sees the data set divided into approximately equal parts. The training set receives parts while the test set receives the remaining part. Under complex sampling -fold cross-validation will be applied to each stratum by dividing the PSUs into approximately equal parts. Refer to the stratified two-stage cluster design described before where the th stratum has PSUs. Then, PSUs are retained as a test set while the remaining PSUs become the training set. The sampling weights associated with the units within the training set have to be adjusted to compensate for the deleted PSUs such that the sum of the sampling weights still equals the correct population total.

In Rao et al. (1992) a sampling weight adjustment is proposed for the delete-1 jackknife method applied to CS data whereby the sampling weights of the remaining units, after some PSU has been deleted, are adjusted upwards by a factor . Here is the original number of PSUs in stratum and is the remaining number of PSUs after a single repetition of the delete-1 jackknife method. Following this reasoning, the proposed sampling weight adjustment of the units in the training set, under KCV, is as follows:

where is the number of folds used for the cross-validation. Let denote the original sampling weight associated with the th unit in the th PSU in stratum . The factor is used to adjust upwards, i.e. , , to compensate for the units in the PSUs removed the test set. These new weights are now used when fitting an SWLS model to the training set after which the fitted model is used to predict the test set responses.

Consider the th part as the test set and let the th response of the th PSU in the test set of stratum be denoted by while the predicted response is denoted by . Consider the sampling weights of the units in the test set of which the sum will no longer equal the intended population total and as such need to be adjusted. In this paper it is argued that, since the training set weights have been adjusted upwards to compensate for the deletion of the test set units, the data in the test set are simply new out-of-sample covariates for which a response must be predicted using the fitted model. Hence, the PE for the th test set of stratum will be calculated as

,

where is the number of SSUs in each PSU in the th test set.

This process is repeated for each of the parts into which the PSUs in stratum have been divided resulting in estimated PEs, , in each stratum. The estimated PE of stratum is thus calculated as the average of the estimated PEs,

and the overall estimated PE is calculated as

,

where is the population number of PSUs in stratum and is the total number of PSUs in the population, i.e. .

The choice of in CV is important and, as in the case of i.i.d. data, there is a trade-off between bias and variance in this choice. In the simulation study this trade-off will also be used for choosing .

Remark: Alternatively, one could argue that the training and test sets could be viewed as two independent samples from the same population and as such the sampling weights in both sets need to be adjusted. However, in this paper the former argument will be followed.

**3. Methodology**

*3.1. Data description*

The dataset that will be used in this analysis and that will act as surrogate population is the 2005/2006 Income and Expenditure Survey (IES) of Statistics South Africa. The intention of the IES is to examine income and expenditure in South Africa and in this research it will be used to model personal income based on a selection of covariates.

A number of adjustments were made to the original 2005/2006 IES such that a “clean” dataset could be obtained which then became the surrogate population used in this simulation study. In a nutshell, the only records that were retained are those for which an age of at least 21 and no older than 65 as well as a positive income was captured. This decision was made such that the surrogate population contains persons of working age while keeping in mind that those at least 21 years old include persons that have completed their bachelor’s degrees as well as those that either did not complete school or that did not continue with a post-school education. The covariates identified from the IES for the modelling of personal income, are: age, ; gender (1 = male, 2 = female), ; race (1 = black, 2 = coloured, 3 = indian/asian, 4 = white); and education level (coded from 0 to 26).

The “Black” category of the race variable was considered the reference category, since it had the largest proportion of observations in the surrogate population, and dummy variables , , and were formed for the remaining three race categories. The education level variable was re-grouped into 7 categories, i.e. no, some primary, complete primary, early high school, non-completed high school, completed high school, and post high school education, of which “no education” was considered the reference category. Dummy variables were formed for the remaining six education level categories.

These predictors comprise the main effects of the income model to which first-order interactions between gender, race and education level, were added. Hence, the IES linear model is given by

*3.2. Simulation study*

Determining which of the PE estimation methods perform “best” requires a comparison of the obtained estimates of PE to the “true” PE. Since the “true” PE is unknown it also needs to be estimated. For this purpose the surrogate population will be considered as the population from which the “truth” can be deduced. Hence, the simulation study for the evaluation of the SWLS model PE consists of two phases: the calculation of the “true” PE; and the comparison of the PE estimation methods to the “true” PE through the evaluation of diagnostic measures.

To determine the “true” PE it was recommended by Molinaro et al. (2005) that a number of samples be selected from the population and that each of these samples be considered a learning set while all observations in the population but not in the learning set form the test set. For this purpose, and also for the comparison of the estimated PEs to the “truth”, samples from the surrogate population were selected. Each sample followed a stratified two-stage cluster design with the nine provinces of South Africa as strata and enumerated areas (EAs), the smallest geographical area into which the country has been divided for survey purposes, as PSUs. The surrogate population consists of PSUs across the 9 strata. Although no clear rule exists as to what the size of learning sets should be, it is recommended that the learning set contain more observations than the test set. Thus, each sample contains between 50% and 60% of these PSUs across the 9 strata. In each selected PSU, four households (HH) were selected and one person per HH was included in the final sample.

At each sampling stage, equal probability sampling was used in the hope that large weight variability would be achieved such that the effect of weight trimming on inference precision could be observed. Differential non-response was also simulated in the design to evaluate the weighting procedures under non-perfect circumstances which is generally found in practice.

Consider the first phase where the “true” PE is estimated and let the replicate samples denote the learning sets. Let the population number of PSUs be denoted by where , and is the number of PSUs in stratum , . Consider the th replicate with PSUs, where is the number of PSUs in stratum , and let it denote the learning set on which the SWLS model is fitted. It should be pointed out that, since the replicate samples have been selected based on a CS design, an SWLS model is fitted to the learning set. The test set thus consists of the remaining PSUs to be predicted by the fitted SWLS model. Consider the th stratum in the test set with PSUs and SSUs in the th PSU, . The “true” stratum PE, denoted by , is then calculated as

where . Finally, the “true” PE is calculated as

.

Note that the sampling weights are not used in the calculation of . It is important to use the sampling weights when fitting a linear model to the learning set since the learning set is a complex sample from the population. However, the test set contains the remainder of the population units that are not included in the learning set. Thus, no sampling weights are in question when calculating from the test set.

This is repeated for all replicate samples resulting in estimates of the “true” PE, . The overall estimate of the “true” PE can thus be calculated as the average of the R estimated PEs.

Alternatively, as described in Molinaro et. al (2005), the estimates of the “true” PE can be seen as individual PEs, one for each replicate sample. Both approaches to the estimation of will be considered. Let the first approach be referred to as the Luus approach while the second approach is referred to as the Molinaro approach. Since the results based on the Molinaro approach are so similar to the results based on the Luus approach they will not be shown here.

The replicates have a second purpose in the simulation study, namely as a sample from which the PE can be estimated by the cross-validation methods discussed in section 2. Diagnostic measures, namely bias and mean squared error (MSE), for estimators obtained from three types of weighting will be compared where the estimates were obtained using both the untrimmed and the trimmed weights. These are: design weight (Design); raking ratio, integrated weighting based on person auxiliary variables (); and raking ratio, integrated weighting based on person and household auxiliary variables (). The person level auxiliary variables (pp) used in the construction of the integrated weights, are: province (9 categories); gender (2 categories); race (4 categories); and age (4 categories). The person and household level auxiliary variables (ph): all four person level auxiliary variables; area (2 categories); dwelling type (2 categories); and household size (3 categories).

The weight trimming methods used will be the 1.5IQR threshold and the Hill threshold introduced by Luus (2016). For information on these trimming methods the reader is invited to consult the referenced literature.

For each PE estimation method and for each type of weighting, both untrimmed and trimmed, the diagnostic measures are compared and a subset of the results is presented in the next section.

**4. Discussion of results**

*4.1. Choice of K*

The choice of for cross-validation is of importance and, generally, it is chosen based on a bias-variance trade-off. For the i.i.d case it has been found that when or the estimated PEs have neither a high bias nor a high variance (James, et al., 2013). Is the same true under CS data? Table 1 shows the median estimated PE as well as the standard deviation for different values of K, namely 2, 5, 10, 15, 20 and n (LOOCV), obtained when fitting the SWLS using the untrimmed design weight and benchmarked weights.

Table 1: Median and standard deviation of estimated PE for different K

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STAT** |  | **R=1:10** | | | **R=1:20** | | | **R=1:30** | | | **R=1:40** | | | **R=1:50** | | |
| **Median** | **K** | **Design** | **RRpp** | **RRph** | **Design** | **RRpp** | **RRph** | **Design** | **RRpp** | **RRph** | **Design** | **RRpp** | **RRph** | **Design** | **RRpp** | **RRph** |
|  | **2** | 13.88 | 13.34 | 16.68 | 15.23 | 13.44 | 14.15 | 14.17 | 12.47 | 13.03 | 13.57 | 11.55 | 13.03 | 13.13 | 12.18 | 13.37 |
|  | **5** | 10.11 | 9.30 | 11.57 | 10.45 | 11.72 | 11.57 | 9.54 | 9.47 | 10.23 | 9.25 | 9.47 | 9.63 | 9.18 | 9.55 | 9.63 |
|  | **10** | 9.69 | 11.26 | 10.09 | 10.85 | 11.55 | 11.04 | 10.00 | 10.53 | 10.09 | 9.42 | 10.05 | 9.04 | 9.10 | 10.05 | 9.11 |
|  | **15** | 10.00 | 9.89 | 10.04 | 10.77 | 10.88 | 11.04 | 9.80 | 9.30 | 10.06 | **9.37** | **8.88** | **9.25** | **9.63** | **9.06** | **9.67** |
|  | **20** | 9.68 | 10.07 | 10.19 | 10.72 | 10.69 | 10.59 | 9.56 | 9.49 | 9.82 | **8.85** | **8.85** | **9.25** | **8.96** | **8.95** | **9.15** |
|  | **n** | 9.34 | 9.59 | 9.60 | 10.76 | 10.85 | 10.80 | 9.34 | 9.45 | 9.38 | **9.09** | **9.03** | **9.17** | **9.14** | **9.04** | **9.14** |
| **Std Dev** | **K** | **Design** | **RRpp** | **RRph** | **Design** | **RRpp** | **RRph** | **Design** | **RRpp** | **RRph** | **Design** | **RRpp** | **RRph** | **Design** | **RRpp** | **RRph** |
|  | **2** | 2.95 | 2.59 | 5.37 | 53.54 | 5.73 | 7.31 | 44.30 | 5.24 | 7.31 | 41.55 | 4.91 | 8.20 | 72.29 | 6.66 | 15.59 |
|  | **5** | 2.23 | 1.97 | 3.29 | 4.29 | 5.49 | 4.55 | 4.33 | 5.18 | 4.53 | 4.25 | 5.04 | 4.48 | 8.08 | 23.48 | 4.44 |
|  | **10** | 2.30 | 2.12 | 2.13 | 4.30 | 4.37 | 4.56 | 4.30 | 4.56 | 4.44 | 3.93 | 4.26 | 4.05 | 4.29 | 4.16 | 4.45 |
|  | **15** | 2.80 | 2.09 | 2.24 | 4.22 | 3.47 | 4.24 | 4.14 | 3.68 | 4.17 | **3.87** | **3.44** | **3.91** | **3.96** | **3.73** | **3.88** |
|  | **20** | 1.84 | 2.18 | 1.96 | 4.12 | 4.47 | 3.65 | 4.09 | 4.31 | 3.72 | **3.74** | **3.93** | **3.40** | **3.84** | **3.93** | **3.64** |
|  | **n** | 2.10 | 2.15 | 2.10 | 3.85 | 3.84 | 3.73 | 3.93 | 3.93 | 3.84 | **3.62** | **3.62** | **3.53** | **3.91** | **3.82** | **3.75** |

It can be seen that the median prediction error as well as the standard deviation stabilize at approximately samples and or . Thus, results will be presented for and for LOOCV based on samples.

*4.2. Discussion of results*

Let the estimate of the PE obtained from the th replicate sample be denoted as . The “true” bias and MSE of , following the Luus approach to obtain the “true” test PE, are approximated by

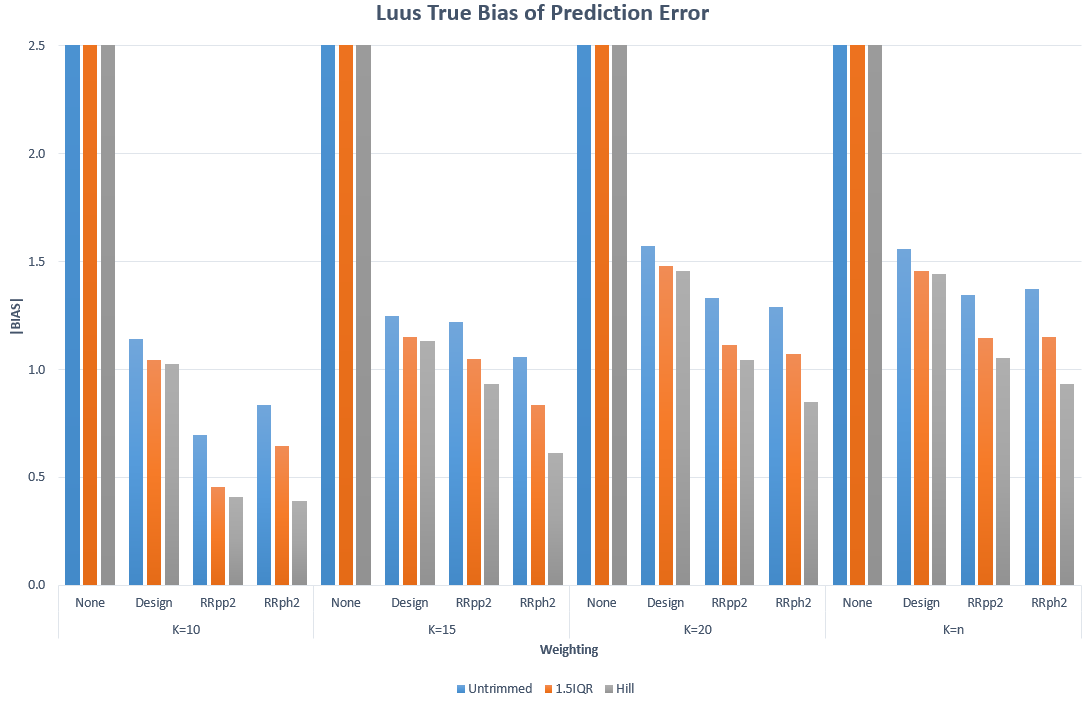
,

and

.

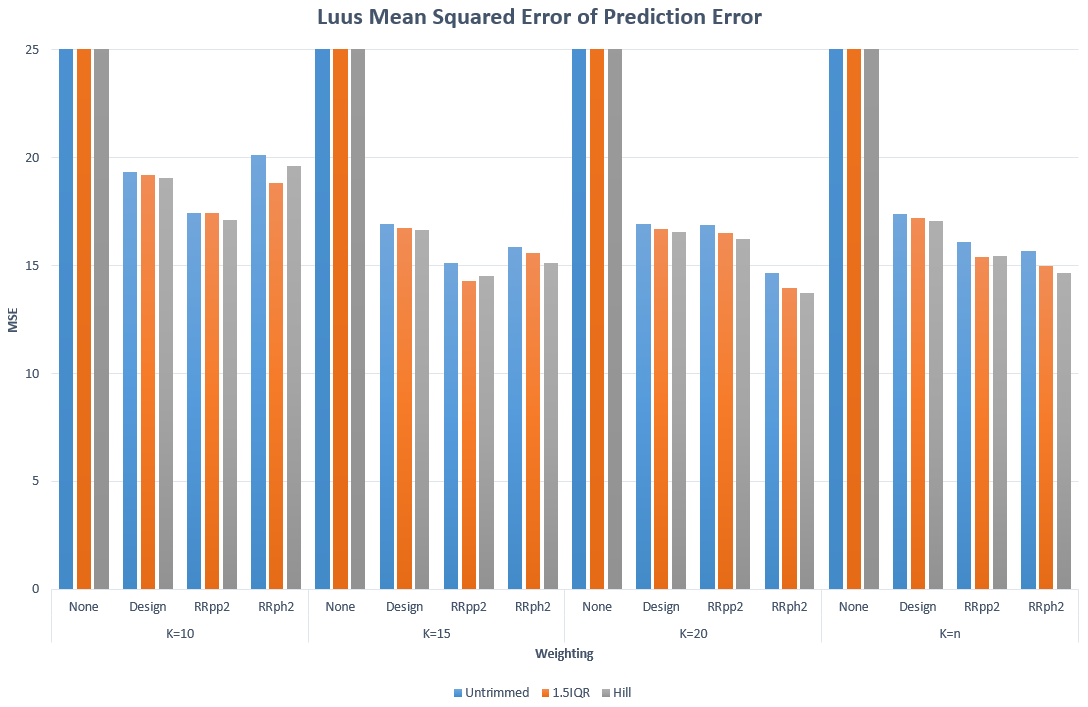
The results based on the Luus approach to “true” PE are shown in Figure 1and Figure 2.

Figure 1: "True" bias of Prediction Error Estimator



The bias appears to reach a minimum when K is equal to 10. Furthermore, the minimum bias is obtained when using the Hill trimmed person and household benchmarked weights ().

Figure 2: "True" MSE of Prediction Error Estimator



In Figure 2 it is seen that the “true” MSE achieves a minimum when and the Hill trimmed person and household benchmarked weights () are used for the modelling.

**5. Conclusion**

The model often used to define the relationship between the response and the predictors is assumed to be a linear model. When modelling a linear relationship between the response and predictors obtained from CS sampling, the SWLS model is employed. Since modelling is often applied with the aim to predict a future response, it is important to be able to evaluate how well the model performs in this regard. Cross-validation has long been used, in the i.i.d. case, for the estimation of a model’s prediction error, but is fairly unfamiliar in the CS sampling case. This paper developed -fold cross-validation for the prediction error estimation of the SWLS model. A simulation study, based on the IES 2005/2006 survey, was used to, on the one hand, determine the optimal size of in the SWLS case, and, on the other, evaluate the use of cross-validation to estimate the PE of the SWLS model under different sampling weights, both trimmed and untrimmed. It was found that at least splits are required to adequately capture the variance structure of the CS data. Furthermore, the Hill trimmed person and household level benchmarked weights resulted in PEs with the smallest bias and MSE.

# **6.** **References**

Heeringa, S., West, G. & Berglund, P., 2010. *Applied survey data analysis.* s.l.:Taylor and Francis Group.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. *An introduction to statistical learning with applications in R.* s.l.:Springer.

Lohr, S., 2010. *Sampling: Design and Analysis.* s.l.:Brooks/Cole.

Luus, R., 2016. *Statistical inference of the multiple regression analysis of complex survey data.* Stellenbosch: Stellenbosch University.

Molinaro, A., Simon, R. & Pfeiffer, R., 2005. Prediction error estimation: a comparison of resampling methods. *Bioinformatics,* 21(15), pp. 3301-3307.

Neethling, A. & Galpin, J., 2006. Weighting of Household Survey Data: A Comparison of Various Calibration Integrated and Cosmetic Estimators. *The South African Statistical Journal,* pp. 123-150.

Rao, J., Wu, C. & Yue, K., 1992. Some recent work on resampling methods for complex surveys. *Survey Methodology,* 18(2), pp. 209-217.