South African Statistical Journal
Vol. 54, No. 2, 187–205
https://doi.org/10.37920/sasj.2020.54.2.5

PREDICTION ERROR ESTIMATION OF THE SURVEY-WEIGHTED LEAST SQUARES MODEL UNDER COMPLEX SAMPLING

Retha Luus[1]
Department of Statistics and Population Studies, University of the Western Cape,
Cape Town, South Africa
e-mail: *rluus@uwc.ac.za*

Ariane Neethling
Department of Statistics and Actuarial Science, University of the Free State,
Bloemfontein, South Africa

Tertius de Wet
Department of Statistics and Actuarial Science, Stellenbosch University,
Stellenbosch, South Africa

Linear modelling with the objective to predict a future response is ubiquitous in statistical analysis. Methods such as cross-validation and the bootstrap are well known for estimating the predictive performance of a model fitted to i.i.d. data. However, many large-scale surveys make use of a complex sampling design where the data are no longer i.i.d. and sampling weights are assigned to each observation to account for this. This paper shows how the cross-validation and bootstrap methods need to be adapted to evaluate the predictive performance of the survey-weighted least squares model. The investigation of the performance of the different prediction error estimation methods is 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 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, are introduced. It was found that the bootstrap estimator of prediction error achieved lower mean squared error while the K-fold cross-validation estimator achieved lower bias. Further improvement was observed using the 1.5IQR and Hill truncated sampling weights.

Key words: Bootstrap, Calibration and integrated weighting, Cross-validation, Prediction error, Survey-weighted least squares, Trimming.

1. Introduction

One of the objectives 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 set of covariate values. As a measure of performance of the linear model in modelling the data, the prediction error (PE) is often used, 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

[1] Corresponding author.

MSC2010 subject classifications. 62D05.

model's predictive capability. However, future observations are unknown and as such not available for assessment of the linear model (James et al., 2013).

Cross-validation (CV) is a well-known approach, under independent and identically distributed (i.i.d.) data, for the construction of out-of-sample data sets by splitting the sample data into training and test sets. The model is developed using the training sets after which it is used to predict the responses in the test sets. The data in the test sets do not form part of the data used for the development of the statistical models and thus the test sets are the out-of-sample data sets. At a minimum the split is done once, viz. validation set (VS) cross-validation, or K times, viz. K-fold cross-validation. When $K = n$, viz. the test set contains one observation unit at a time, then it is called leave-one-out cross-validation (LOOCV) (James et al., 2013).

Consider a finite population of size N with an N-vector of responses, $\underline{y}_U$, and p predictor variables, $\underline{x}_1, \ldots, \underline{x}_p$, and let x_{ij} represent the value of the ith population element of the jth predictor variable for $i = 1, \ldots, N$ and $j = 1, \ldots, p$. 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,

$$\underline{y}_U = \underline{X}_U \underline{\beta} + \underline{\varepsilon},$$

where $\underline{X}_U$ is an $N \times p$ matrix of population predictors, $\underline{\beta}$ is a p-dimensional vector of unknown regression coefficients and each element of $\underline{\varepsilon}$ has a $N(0, \sigma^2)$ distribution (Lohr, 2010).

This paper considers the estimation of the prediction error of the linear model applied to data obtained through complex sampling (CS). The cross-validation methods, i.e. VS, K-fold and LOOCV, are not well known in complex sampling research, and introducing and evaluating them is the contribution of this paper. One can also estimate the model's prediction error using the bootstrap. As such, and as an alternative to the CV methods, the bootstrap PE method is discussed and then extended for use under CS. The performance of the bootstrap PE method as a way of evaluating a model's predictive capability will be compared to that of the CV methods.

Consider a sample selected through a stratified two-stage cluster design whereby a population has been stratified into H strata and within stratum h there are N_h primary sampling units (PSUs) of which n_h PSUs are selected. Let the hjth selected PSU contain N_{hj} secondary sampling units (SSUs) and suppose a sample of n_{hj} SSUs is selected from this PSU, $j = 1, \ldots, N_h$ and $h = 1, \ldots, H$. Each unit in this CS is assigned a design weight, $d_{hji}, i = 1, \ldots, n_{hj}, j = 1, \ldots, n_h, h = 1, \ldots, H$, calculated as the inverse of the inclusion probability of the hjith unit. It is a number 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 (Deville and Särndal, 1992; Neethling and Galpin, 2006; Lohr, 2010). After these weight development stages have been completed the sampling weights, $w_{hji}, i = 1, \ldots, n_{hj}, j = 1, \ldots, n_h, h = 1, \ldots, H$, are obtained. The following is a summary of the integrated weighting technique used to obtain the sampling weights of a household-based survey.

Define $\underline{a}_k$ as the M-vector of M auxiliary variable values for person k in household hh. For the person level auxiliary variables, the entries $\underline{a}_k$ are the proportion of members in the household to which person k belongs, that have the corresponding auxiliary characteristics (e.g. proportion of males in the household, proportion of females in the household, etc.). For a household level auxiliary variable, e.g. dwelling type, define a new variable for each category of the household auxiliary variable under consideration, e.g. urban and rural. Then the value for the category to which the household belongs is simply the inverse of the household size, and 0 for all other categories of the household variable (Neethling and Galpin, 2006).

The integrated weight associated with person k, w_k, is formed by minimising a given distance function between the design weight and w_k subject to a set of constraints. Different distance functions can be used in practice for minimising the distance between the design weight and w_k. In this paper the linear distance and the exponential distance, or raking ratio, will be used (Neethling and 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, leading to

$$\underline{\hat{\beta}}_{SWLS} = \left(\underline{X}'\underline{W}\underline{X}\right)^{-1} \underline{X}'\underline{W}\underline{y},$$

where $\underline{X}$ is an $n \times p$ matrix of predictors, $\underline{W}$ is the $n \times n$ diagonal sampling weight matrix, with $n = \sum_{h=1}^{H} \sum_{j=1}^{n_h} n_{hj}$, and $\underline{y}$ is the n-vector of responses. Note that, even though $\underline{\hat{\beta}}_{SWLS}$ is similar to the estimated model coefficients obtained under weighted least squares (WLS) regression, the variance of $\underline{\hat{\beta}}_{SWLS}$ is not the same as the variance under WLS. Herein lies the importance of using SWLS when linearly modelling CS data, since standard errors, confidence intervals and hypothesis tests will be wrong otherwise (Lohr, 2010).

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 Sections 4 and 5.

The purpose of this paper is to estimate the prediction error of the SWLS model using cross-validation and the bootstrap. The next section introduces cross-validation and its application to evaluate the predictions of the linear model with i.i.d. errors. The general CV methods are then developed for SWLS model evaluation. The bootstrap approach to the estimation of PE is presented in Section 3. Section 4 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 5 the results obtained from this analysis are presented and discussed and, finally, conclusions and areas for further research are given in Section 6.

2. Prediction error estimation using cross-validation

Consider a simple random sample (SRS) of n observations where each observation is associated with a p-vector of measured covariates, $\underline{x}$, and a continuous response, y, with an unknown distribution, P. One of the aims of statistical modelling is the construction of a rule by which to predict a future unobserved outcome, say y_0, at its associated covariate value $\underline{x}_0$. In this paper the linear model is assumed due to it most often being used in applied statistics to model the response as a function of the covariates. Thus, if $\hat{y}$ denotes the predicted outcome, $\underline{x}$ the associated vector of covariates and $\underline{\hat{\beta}}$ the vector of estimated model coefficients, then, under the linear model,

$$\hat{y} = \underline{x}'\underline{\hat{\beta}}.$$

To evaluate the performance of a prediction rule one can make use of loss functions and most commonly the squared error loss, given by

$$L(y) = \left(y - \underline{x}'\underline{\hat{\beta}}\right)^2,$$

is used. For the purpose of prediction rule evaluation, define the expected loss

$$E(L(y)) = E\left(y - \underline{x}'\underline{\hat{\beta}}\right)^2.$$

In this paper the aim with the evaluation of the prediction rule is to determine how well the linear model predicts an out-of-sample response. Hence, one is interested in estimating the generalisation error or prediction error (PE) (Molinaro et al., 2005; Hastie et al., 2009).

In an ideal world an independent dataset will be available for the purpose of model selection and to estimate the PE, but in reality the observed data are all one has available. Estimating the PE using the observed data gives the apparent error, $\widehat{PE}^{Apparent}$,

$$\widehat{PE}^{Apparent} = \hat{E}(L(y)) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \underline{x}_i'\underline{\hat{\beta}}\right)^2.$$

When a dataset is used to construct a prediction rule, the fitting method used to construct the rule adapts to the data to which it is fitted. Hence, using the same data to construct the rule and evaluate its performance, i.e. using the apparent error to estimate the test error, will lead to an estimated PE that is too optimistic (Molinaro et al., 2005; Hastie et al., 2009).

To address the problem of a biased PE estimate, CV methods have been utilized to construct artificial extra-samples to be used as "new" observations to be predicted by the constructed prediction rule. Although well-known for the SRS case, cross-validation as PE estimation method is not as well-known in the CS case and as such the CV methods will firstly be described for the SRS case in Section 2.1 and then developed for the CS case in Section 2.2.

2.1 Cross-validation under simple random sampling

Considered to be the simplest and most widely used method for estimating PE, CV splits the data into a set on which the model is fitted and a set on which the fitted model is tested. At a minimum the data are split once into two parts. This is called the validation set (VS) approach. This method, however, does not necessarily capture the data structure adequately resulting in highly variable PE estimates when the method is carried out repeatedly. Also, the validation set approach tends to overestimate the true PE (James et al., 2013).

As remedy for the drawbacks of the validation set approach the leave-one-out cross-validation (LOOCV) method was proposed where the data are split into n parts, one part for each observation.

Let the part containing the ith observation be the test sample and let the remaining $n - 1$ parts be used to fit the linear model. Let the predicted value of the ith observation be $\hat{y}_{i(i)}$. The error in predicting the ith observation is calculated as

$$\widehat{PE}^{(i)} = \left(y_i - \hat{y}_{i(i)}\right)^2 .$$

This results in n estimates of PE, $\{\widehat{PE}^{(i)}\}$ for $i = 1, \ldots, n$, and finally the LOOCV estimated PE is calculated as (James et al., 2013)

$$\widehat{PE}^{LOOCV} = \frac{1}{n} \sum_i \widehat{PE}^{(i)}.$$

The LOOCV estimate of PE is less biased and less variable than the validation set estimated PE. Also, it tends not to overestimate the true PE. However, if the data set is quite large or if a complex statistical model is being evaluated, then the LOOCV method can be computationally expensive (James et al., 2013).

To improve on the computation time while retaining the advantages of the LOOCV the K-fold cross-validation (KCV) method was developed. Here the data are split into K parts of approximately equal size. $K - 1$ parts are used to fit the model while the remaining part is used for testing the fitted model. Suppose the kth part is retained as the test sample and the remaining $K - 1$ parts are used as a training sample on which the linear model is fitted. Suppose the kth part contains approximately $n_k = \frac{n}{K}$ observations. Let the predicted value of the ith observation in the kth test set be defined as $\hat{y}_{i(k)}$ where the subscript (k) is used to emphasise that the kth part is used as the test set. The estimated PE in this case is calculated as

$$\widehat{PE}^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} \left(y_{i(k)} - \hat{y}_{i(k)}\right)^2 ,$$

where $\widehat{PE}^{(k)}$ denotes the estimated PE of the kth test sample. This procedure is repeated for all K parts resulting in K estimates of PE, viz. $\widehat{PE}^{(1)}, \ldots, \widehat{PE}^{(K)}$. The final K-fold CV estimated PE is obtained as the average of the K estimated PEs (James et al., 2013),

$$\widehat{PE}^{KCV} = \frac{1}{K} \sum_k \widehat{PE}^{(k)}.$$

In K-fold CV both the proportion of observations in the test set and the number of estimates to average can affect the error estimate. When increasing K the proportion of observations in the test set decreases while the proportion in the training set increases. This will cause a decrease in bias. Furthermore, a large number of estimates to average may also decrease the bias (Molinaro et al., 2005). However, increasing K significantly could also bring about an increase in the variability of the estimated PE. Careful consideration of this bias-variance trade-off determined that using $K = 5$ or 10 yields estimated PEs that are neither highly biased nor highly variable (James et al., 2013).

2.2 Cross-validation under complex sampling

Now consider the application of these PE estimation methods to the stratified two-stage cluster design described before. 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 K-fold cross-validation sees the data set divided into K approximately equal parts. The training set receives $K-1$ parts while the test set receives the remaining part. Under complex sampling K-fold cross-validation will be applied to each stratum by dividing the PSUs into K approximately equal parts. Consider the hth stratum with n_h PSUs. Then, $\widetilde{n}_h = \frac{n_h}{K}$ PSUs are retained as a test set while the remaining $n_h - \widetilde{n}_h$ 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 $\frac{n_h}{n_h-1}$. Here n_h is the original number of PSUs in stratum h and $n_h - 1$ 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:

$$\frac{n_h}{n_h - \widetilde{n}_h} = \frac{n_h}{n_h - \frac{n_h}{K}} = \frac{n_h}{n_h\left(1 - \frac{1}{K}\right)} = \frac{K}{K-1},$$

where K is the number of folds used for the cross-validation. Let w_{hji} denote the original sampling weight associated with the ith unit in the jth PSU in stratum h. The factor $\frac{K}{K-1}$ is used to adjust w_{hji} upwards, i.e. $w_{hji} \cdot \frac{K}{K-1}, i = 1, \ldots, n_{hj}, j = 1, \ldots, (n_h - \widetilde{n}_h)$, to compensate for the units in the PSUs removed from 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 kth part as the test set and let the ith response of the jth PSU in the test set of stratum h be denoted by $y_{hji}^{(k)}$ while the predicted response is denoted by $\hat{y}_{hji}^{(k)}$. 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 kth test set of stratum h will be calculated as

$$\widehat{PE}_h^{(k)} = \frac{1}{\widetilde{n}_h} \sum_{j=1}^{\widetilde{n}_h} \frac{1}{n_{hj}} \sum_{i=1}^{n_{hj}} \left(y_{hji}^{(k)} - \hat{y}_{hji}^{(k)}\right)^2,$$

where n_{hj} is the number of SSUs in each PSU in the kth test set.

This process is repeated for each of the K parts into which the PSUs in stratum h have been divided resulting in K estimated PEs, $\widehat{PE}_h^{(1)}, \ldots, \widehat{PE}_h^{(K)}$, in each stratum. The estimated PE of stratum h is thus calculated as the average of the K estimated PEs,

$$\widehat{PE}_h = \frac{1}{K} \sum_{k=1}^{K} \widehat{PE}_h^{(k)}, h = 1, \ldots, H,$$

and the overall estimated PE is calculated as

$$\widehat{PE}_{SWLS}^{KCV} = \frac{\sum_{h=1}^{H} N_h \widehat{PE}_h}{N},$$

where N_h is the population number of PSUs in stratum h and N is the total number of PSUs in the population, i.e. $N = \sum_h N_h$.

The validation set approach and the leave-one-out cross-validation are special cases of the KCV. Consider again stratum h with n_h PSUs. Under the validation set approach, $K = 2$, while under the leave-one-out cross-validation, $K = n_h$. Using the sampling weight adjustment discussed before, the sampling weights of the units in the training sets, created by these methods, are also adjusted upwards to compensate for the PSUs that have been removed to the test set.

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. Prediction error estimation using the bootstrap

This section presents two bootstrap methods for prediction error (PE) estimation as alternatives to the well-known cross-validation methods.

3.1 Bootstrap estimator of prediction error under simple random sampling

Consider an SRS of size n with data $(y_i, \underline{x}_i')$, $i = 1, \ldots, n$, to which a linear model is fitted. The model can be evaluated by estimating the response of the sample from which the model was obtained and calculating its PE. The PE calculated in this regard is called the apparent error rate,

$$\widehat{PE}^{Apparent} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

where y_i is the observed and $\hat{y}_i$ the estimated response of the ith observation (Efron and Tibshirani, 1998).

Efron and Tibshirani (1998) describe two approaches to bootstrap regression, namely the bootstrapping residuals approach and the bootstrapping pairs approach. This paper considers the latter approach while the former will be considered under further research.

The bootstrapping pairs approach proceeds by generating a with-replacement bootstrap sample of size n, $\left(y_i^*, \underline{x}_i^{*\prime}\right)$, from $\left(y_i, \underline{x}_i'\right)$ and then the linear model is fitted to this bootstrap sample. This fitted model is firstly used to predict the response of the observed sample, $\underline{\widetilde{y}} = \underline{X}\underline{\widehat{\beta}}^*$, where $\underline{X}$ is the matrix of covariates of the original sample, $\underline{\widehat{\beta}}^* = \left(\underline{X}^{*\prime}\underline{X}^*\right)^{-1} \underline{X}^{*\prime}\underline{y}^*$ is the bootstrap estimator, $\underline{X}^*$ is the $n \times p$ matrix of bootstrap predictors and $\underline{y}^*$ is the n-vector of bootstrap responses. The predicted responses are now used to obtain an estimate of the PE,

$$\widehat{PE}^{B_1} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widetilde{y}_i)^2,$$

where the superscript B_1 is used to label the above PE as the PE calculated from the responses of the observed sample predicted by the bootstrap linear model and $\widetilde{y}_i$ denotes the ith predicted response obtained from using the bootstrap linear model to predict the observed sample (Efron and Tibshirani, 1998).

The bootstrap model is used a second time, now to estimate the responses of the bootstrap sample, $\underline{\hat{y}}^* = \underline{X}^*\underline{\widehat{\beta}}^*$, where $\underline{X}^*$ is the bootstrap matrix of covariates. These estimated responses are used to obtain a second PE estimate,

$$\widehat{PE}^{B_2} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^* - \hat{y}_i^*\right)^2,$$

with superscript B_2 used to emphasise that the PE is calculated using the estimated bootstrap responses. Finally, the difference between the two estimated PEs is calculated,

$$\widehat{Diff} = \widehat{PE}^{B_1} - \widehat{PE}^{B_2}.$$

The process outlined above is repeated a large number of times, say B, resulting in B differences, $\{\widehat{Diff}_b\}, b = 1, \ldots, B$, which are used to calculate an optimism,

$$\text{optimism} = \frac{1}{B}\sum_{b=1}^{B}\widehat{Diff}_b,$$

a number that represents the amount by which the apparent error underestimates the true PE (Efron and Tibshirani, 1998). Finally, the bootstrap estimate of PE is obtained as the sum of the apparent error and the optimism,

$$\widehat{PE}^{BS} = \widehat{PE}^{Apparent} + \text{optimism}.$$

3.2 Bootstrap estimator of prediction error under complex sampling

Consider the CS design described before, i.e. a stratified two-stage cluster design. A SWLS is fitted to the observed CS and the model is used to estimate the response of the sample. The estimated responses are used to calculate the apparent PE which, under CS, is given by

$$\widehat{PE}_{SWLS}^{Apparent} = \frac{\sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{i=1}^{n_{hj}} w_{hji}\left(y_{hji} - \hat{y}_{hji}\right)^2}{\sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{i=1}^{n_{hj}} w_{hji}},$$

where y_{hji}, $\hat{y}_{hji}$ and w_{hji} are, respectively, the observed and estimated response and the sampling weight of the ith SSU in the jth PSU of the hth stratum.

Here the sampling weights are used under SWLS regression and in the calculation of the PEs. In the cross-validation case discussed in Section 2 the training and test sets do not overlap, 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 sampling weights are not used when calculating the PE. However, the SWLS model here is fitted to a CS sample and used to estimate the responses of this sample. Thus, the sampling weights are incorporated in the calculation of $\widehat{PE}_{SWLS}^{Apparent}$ to ensure unbiased estimation.

The bootstrap is applied independently per stratum. Now, within each stratum, select a with-replacement sample of m_h PSUs and let these form the bootstrap sample. Let $\underline{y}^* = \{\underline{y}_h^*\}, h = 1, \ldots, H$ denote the responses and $\underline{X}^* = \{\underline{X}_h^*\}, h = 1, \ldots, H$, denote the predictors corresponding to the PSUs in the bootstrap sample.

Due to the with-replacement sampling, the sampling weights have to be adjusted to compensate for PSUs being over-sampled, under-sampled or not sampled at all. Define m^*_{hj} as the number of times the jth PSU is sampled. The bootstrap weights are then calculated as

$$w^*_{hji} = w_{hji}\left[1 - \sqrt{\frac{m_h}{n_h - 1}} + \left(\sqrt{\frac{m_h}{n_h - 1}}\right)\left(\frac{n_h}{n_h - 1}\right)\cdot m^*_{hj}\right],$$

where w_{hji} is the original sampling weight and w^*_{hji} is the bootstrap adjusted sampling weight, $i = 1, \ldots, n_{hj}, j = 1, \ldots, m_h, h = 1, \ldots, H$ (Rust and Rao, 1996). In this paper m_h will be set equal to $n_h - 1$. As mentioned by Rust and Rao (1996), there is considerable practical benefit as well as little, if any, loss in efficiency by doing so. This choice of m_h simplifies the bootstrap sampling weight adjustment to

$$w^*_{hji} = w_{hji}\left[\left(\frac{n_h}{n_h - 1}\right)\cdot m^*_{hj}\right].$$

The bootstrap weights are used in the SWLS model fitted to the bootstrap sample and the bootstrap model is used in the same two ways as discussed in Section 3.1. Firstly, the bootstrap model is used to predict the responses of the observed CS after which the predicted responses are used to calculate the PE,

$$\widehat{PE}^{B_1}_{SWLS} = \frac{\sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{i=1}^{n_{hj}} w_{hji}\left(y_{hji} - \widetilde{y}_{hji}\right)^2}{\sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{i=1}^{n_{hj}} w_{hji}},$$

where $\widetilde{y}_{hji}$ denotes the hjith predicted response obtained from using the bootstrap SWLS model to predict the observed sample. Next the bootstrap model is used to estimate the responses of the bootstrap sample and these are used to calculate a second estimate of PE,

$$\widehat{PE}^{B_2}_{SWLS} = \frac{\sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{i=1}^{n_{hj}} w^*_{hji}\left(y^*_{hji} - \hat{y}^*_{hji}\right)^2}{\sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{i=1}^{n_{hj}} w^*_{hji}},$$

where y^*_{hji}, $\hat{y}^*_{hji}$ and w^*_{hji} are, respectively, the observed and estimated response and the bootstrap sampling weight of the ith SSU in the jth PSU of the hth stratum in the bootstrap sample. Similar reasoning as before for using the sampling weights in the calculation of $\widehat{PE}^{B_1}_{SWLS}$ and $\widehat{PE}^{B_2}_{SWLS}$, is followed.

These two estimates of PE are used to calculate the previously defined difference,

$$\widehat{Diff} = \widehat{PE}^{B_1}_{SWLS} - \widehat{PE}^{B_2}_{SWLS}.$$

The process outlined above is repeated a large number of times, say B, resulting in B differences, $\{\widehat{Diff}_b\}, b = 1, \ldots, B$, which are used to calculate an optimism,

$$\text{optimism}_{SWLS} = \frac{1}{B}\sum_{b=1}^{B}\widehat{Diff}_b.$$

The bootstrap estimate of PE under CS is then calculated as

$$\widehat{PE}^{BS}_{SWLS} = \widehat{PE}^{Apparent}_{SWLS} + \text{optimism}_{SWLS}.$$

4. Methodology

4.1 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, $\underline{y}$, based on a selection of covariates, $\underline{x}_1, \ldots, \underline{x}_p$.

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, X_1;
- gender (1 = male, 2 = female), X_2;

  A dummy variable was constructed for gender and "female" was chosen as the reference category.
- ethnic group (1 = black, 2 = coloured, 3 = indian/asian, 4 = white);

  Ethnic groups 3 and 4 are very small in comparison to ethnic group 1 and were thus combined in an attempt to not end up with empty groups in the simulation study. "Black" was considered the reference category, since it had the largest proportion of observations in the surrogate population. Dummy variables RD_2, and RD_3 were formed for the remaining two race categories.
- education level (coded from 0 to 26).

  This variable was grouped into 7 categories, i.e. no education, 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 and dummy variables $ED_2, ED_3, \ldots, ED_7$ 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

$$\begin{aligned}\underline{y} &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 RD_2 + \beta_4 RD_3 + \beta_5 ED_2 + \beta_6 ED_3 + \beta_7 ED_4 \\ &\quad + \beta_8 ED_5 + \beta_9 ED_6 + \beta_{10} ED_7 + \beta_{11} X_2 RD_2 + \ldots + \beta_{30} RD_3 ED_7 + \underline{\varepsilon}.\end{aligned}$$

4.2 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 training set while all observations in the population but not in the training set form the test set. For this purpose, and also for the comparison of the estimated PEs to the “truth”, $R = 100$ samples were selected from the surrogate population. 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 $N = 2978$ PSUs across the 9 strata. Although no clear rule exists as to what the size of the training sets should be, it is recommended that they do not contain fewer observations than the test set. Thus, each sample contains 50% of the 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 are generally found in practice.

Consider the first phase where the “true” PE is estimated and let the R replicate samples denote the R training sets. Let the population number of PSUs be denoted by N where $N = \sum_{h=1}^{H} N_h$, and N_h is the number of PSUs in stratum h, $h = 1, \ldots, H$. Consider the rth replicate with $n_r = \sum_{h=1}^{H} n_{h_r}$ PSUs, where n_{h_r} is the number of PSUs in stratum h, and let r denote the training 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 training set. The test set thus consists of the remaining $N - n_r$ PSUs to be predicted by the fitted SWLS model. Consider the hth stratum in the test set with $N_h - n_{h_r}$ PSUs and N_{hj} SSUs in the jth PSU, $j = 1, \ldots, N_h - n_{h_r}$. The “true” stratum PE, denoted by $(\widetilde{PE})_{h_r}$, is then calculated as

$$\left(\widetilde{PE}\right)_{h_r} = \frac{1}{N_h - n_{h_r}} \sum_{j=1}^{N_h - n_{h_r}} \frac{1}{N_{hj}} \sum_{i=1}^{N_{hj}} \left(y_{hji} - \hat{y}_{hji}\right)^2,$$

where $h = 1, \ldots, H$. Finally, the “true” PE is calculated as

$$\left(\widetilde{PE}\right)_r = \sum_{h=1}^{H} \frac{N_h}{N} \left(\widetilde{PE}\right)_{h_r}.$$

Note that the sampling weights are not used in the calculation of $(\widetilde{PE})_r$. It is important to use the sampling weights when fitting a linear model to the training set since the training 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 training set. Thus, no sampling weights are in question when calculating $(\widetilde{PE})_r$ from the test set.

This is repeated for all R replicate samples resulting in R estimates of the "true" PE, $\{(\widetilde{PE})_r\}, r = 1, \ldots, R$. The overall estimate of the "true" PE, $\widetilde{PE}$, can thus be calculated as the average of the R estimated PEs.

Alternatively, as described in Molinaro et al. (2005), the R estimates of the "true" PE can be seen as R individual PEs, one for each replicate sample. Both approaches to the estimation of $\widetilde{PE}$ 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.

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 and the bootstrap discussed in Sections 2 and 3. Diagnostic measures, namely bias and mean squared error (MSE), for estimators obtained from five types of weighting will be compared where the estimates were obtained using both the untrimmed and the trimmed weights. These are: design weight (Design); linear calibrated and integrated weighting based on person auxiliary variables (Lin_{pp}); linear calibrated and integrated weighting based on person and household auxiliary variables (Lin_{ph}); raking ratio calibrated and integrated weighting based on person auxiliary variables (RR_{pp}); and raking ratio calibrated and integrated weighting based on person and household auxiliary variables (RR_{ph}). 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 (9 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 4Avg, 1.5IQR, 3.5Med and the Hill. The 1.5IQR and the Hill thresholds were introduced by Luus (2016). The remaining methods are well-known thresholds already in use (Izrael et al., 2009; Valliant et al., 2013). 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.

5. Discussion of results

SAS® was used for sampling from the surrogate population and for the calculation of the different sampling weights while the analyses were performed in R.

5.1 Choice of K

The choice of K for cross-validation is of importance and, for SRS data, it is chosen based on a bias-variance trade-off. Consequently it has been found that when $K = 5$ or $K = 10$ the estimated PEs have neither a high bias nor a high variance (James et al., 2013). Is the same true for CS data? This section presents results to be used to answer this question.

Figure 1 shows the average estimated PE for different values of K, namely 2, 5, 10, 15, 20, n (LOOCV), as well as the bootstrap estimated PE based on $B = 500$ bootstrap samples selected from each sample r, $r = 1, \ldots, 100$, obtained when fitting the SWLS using the untrimmed design weight and benchmarked weights.

In Figure 1 a significant decrease in the average estimated PE is observed as K is increased from 2 to 10. It can be seen that the average prediction error stabilises at approximately $R = 40$ samples and

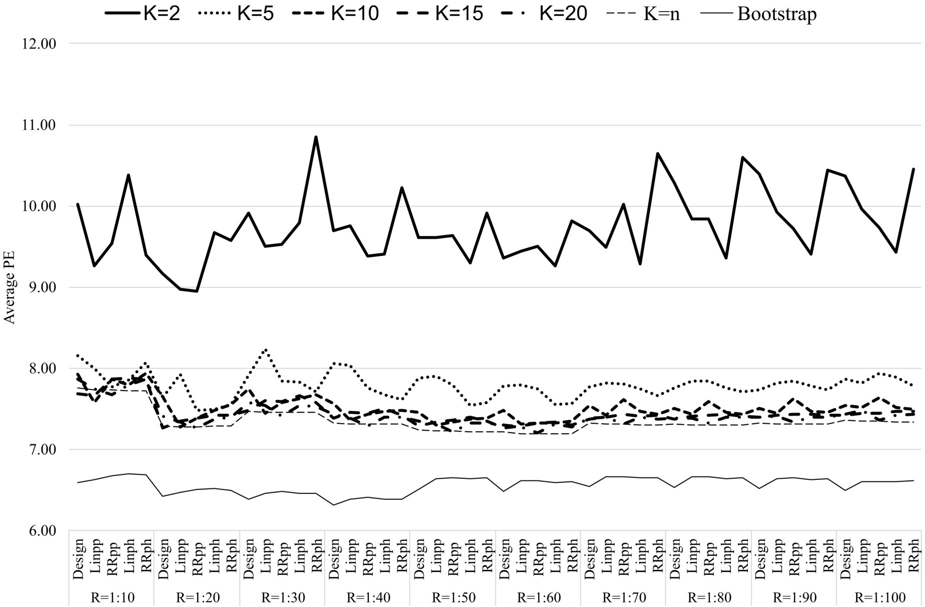

Figure 1. Average estimated PE.

$K \geq 10$. Furthermore, it is observed that the average estimated PE is smallest under the bootstrap PE estimation approach and remains fairly stable across the range of R. A similar trend is observed when considering the median estimated PE shown in Figure 2.

The standard deviations of the estimated PEs are shown in Figure 3. Once again, a significant decrease in standard deviation is observed as K is increased from 2 to 10. The standard deviation of the bootstrap PE estimates is again the smallest, but when $R > 40$ the difference is not as distinguishable. It appears that K should be at least 10 when estimating the PE of the SWLS model.

Thus, results for $K = 10, 15, 20, n$ (LOOCV) and the bootstrap are shown in the next section. Since the linear calibrated and integrated weighting, i.e. Lin_{pp} and Lin_{ph}, respectively, generally did not perform as well as the raking ratio calibrated and integrated weighting, results based on Lin_{pp} and Lin_{ph} will not be shown here.

5.2 Results

Let the estimate of the PE obtained from the rth replicate sample be denoted by $\widehat{PE}_r, r = 1, \ldots, R$. The "true" bias and MSE of $\widehat{PE}$, following the Luus approach to obtain the "true" test PE, are approximated by

$$\left|\text{bias}^L\left(\widehat{PE}\right)\right| = \left|\left(\frac{1}{R}\sum_r \widehat{PE}_r\right) - \widetilde{PE}\right|,$$

and

$$\text{MSE}^L\left(\widehat{PE}\right) = \frac{1}{R}\sum_r \left(\widehat{PE}_r - \widetilde{PE}\right)^2 .$$

The results based on the Luus approach to "true" PE are shown in Figure 4 and Figure 5. When

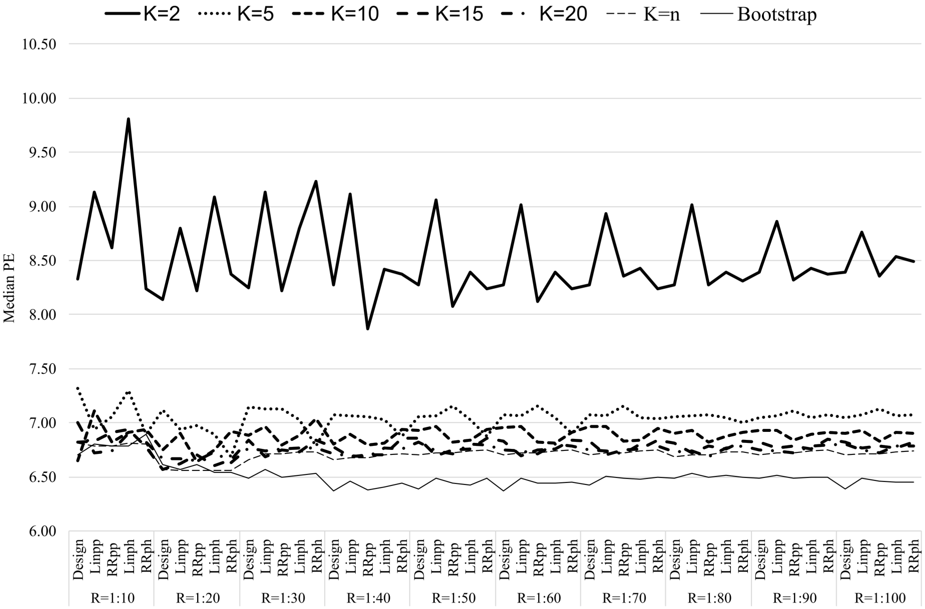

Figure 2. Median estimated PE.

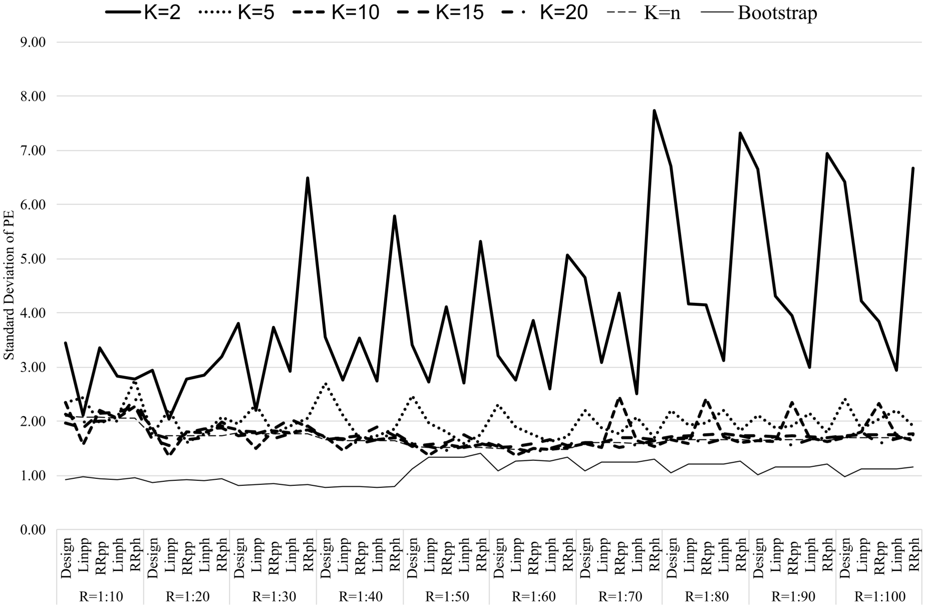

Figure 3. Standard deviation of estimated PE.

following the Molinaro approach the "true" bias and MSE of $\widehat{PE}$ are approximated by

$$\left|\text{bias}^M\left(\widehat{PE}\right)\right| = \left|\frac{1}{R}\sum_r\left(\widehat{PE}_r - \widetilde{PE}_r\right)\right|,$$

and

$$\text{MSE}^M\left(\widehat{PE}\right) = \frac{1}{R}\sum_r\left(\widehat{PE}_r - \widetilde{PE}_r\right)^2.$$

These results are shown in Figure 6 and Figure 7. For each estimation method of PE and for each type of weighting, both untrimmed and trimmed, the diagnostic measures are compared, and a subset of the results are presented below.

The "true" bias of the estimated PE, based on the Luus approach, for $K = 10, 15, 20, n$ (LOOCV) and the bootstrap using the design and benchmarked weights, untrimmed and trimmed, are shown in Figure 4. The bias appears to reach a minimum when K is equal to 10. Furthermore, the minimum bias is obtained when using the 1.5IQR trimmed person benchmarked weights (RR_{pp}). However, bias appears to stabilise when $K \geq 15$ using the Hill trimmed person and household benchmarked weights (RR_{ph}). The "true" bias of the bootstrap estimator of PE is larger than the cross-validation PE estimators. However, the bias is reduced when using the 1.5IQR trimmed RR_{ph} weights. The "true" bias based on the Molinaro approach, shown in Figure 6, achieves the same results.

The "true" MSE (Luus approach), shown in Figure 5, achieves a minimum under the bootstrap PE estimation approach using the 1.5IQR trimmed person benchmarked weights (RR_{pp}). However, the "true" MSE (Luus approach) approximately stabilises when $K \geq 15$ using the Hill trimmed person benchmarked weights (RR_{pp}).

The Molinaro "true" MSE in Figure 7 differs from the result in Figure 5. Here, the "true" MSE is much larger than the "true" MSE in Figure 5. This implies that the variance of the Molinaro "true" PE is much larger than the variance of the Luus "true" PE since the Luus and Molinaro "true" biases are the same. This is not surprising since the Molinaro "true" PE approach is based on a single training set and test set split. This is known to result in more variable PEs than when using more training set test set splits and then averaging over the multiple "true" PEs, as is the case under the Luus approach. However, the minimum is also achieved when the bootstrap PE estimation approach is used based on the 1.5IQR trimmed person benchmarked weights (RR_{pp}).

6. Conclusions and further research

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 unknown in the CS sampling case. In comparison to cross-validation the bootstrap approach to estimating PE is not as well-known, even more so under complex sampling.

This paper extended K-fold cross-validation and the bootstrap to be used 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 K in the SWLS case, and, on the other, evaluate

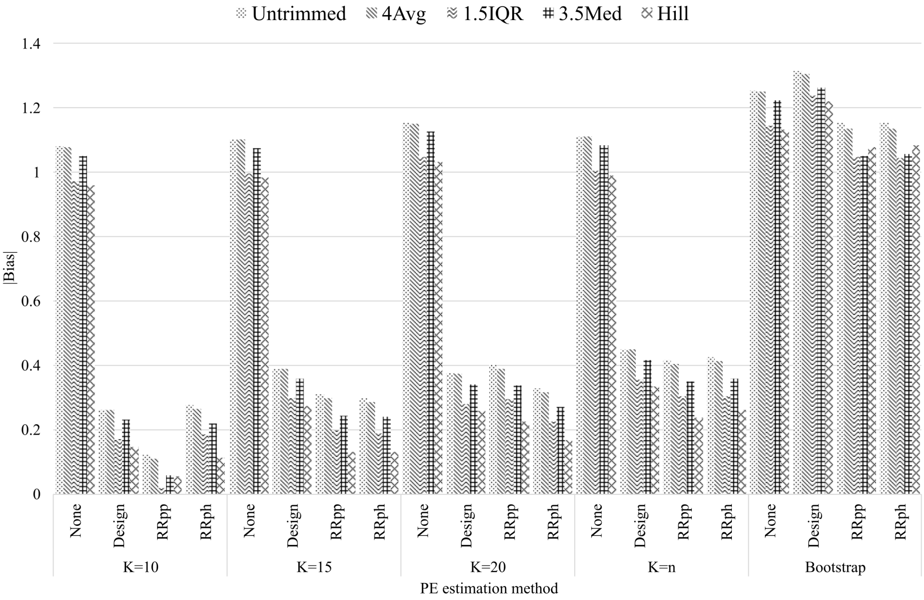

Figure 4. Comparing the “true” bias of different CV estimators and the bootstrap estimator of PE using the Luus approach.

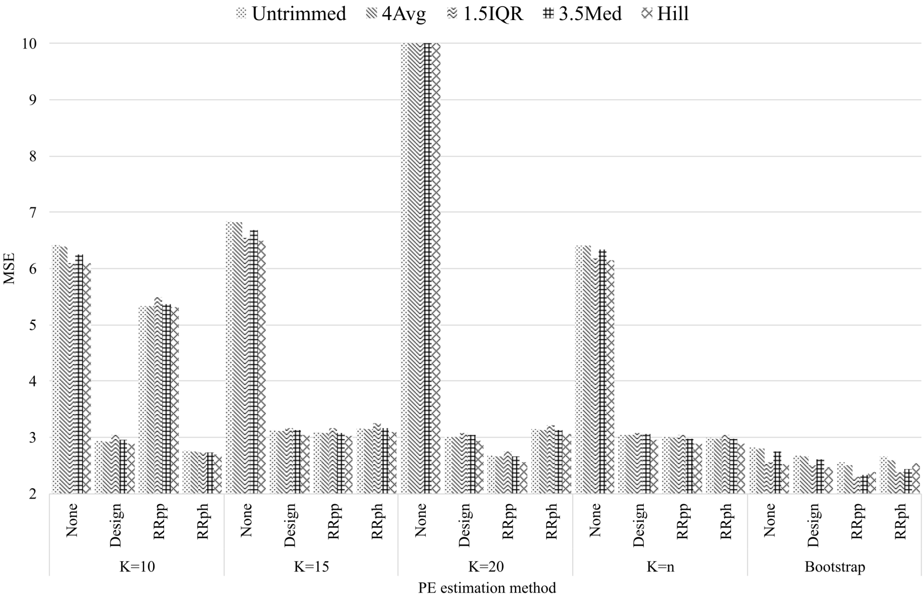

Figure 5. Comparing the “true” MSE of different CV estimators and the bootstrap estimator of PE using the Luus approach.

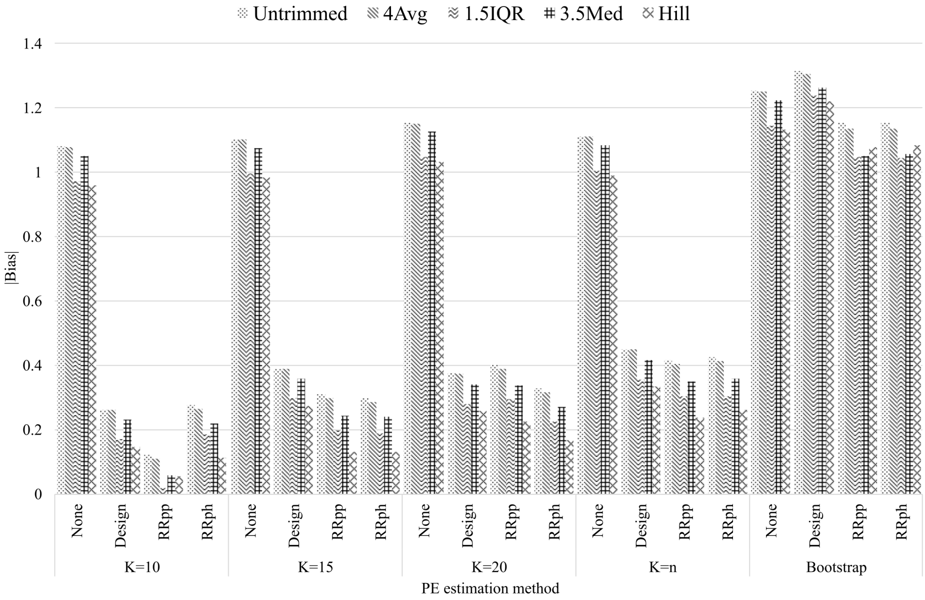

Figure 6. Comparing the “true” bias of different CV estimators and the bootstrap estimator of PE using the Molinaro approach.

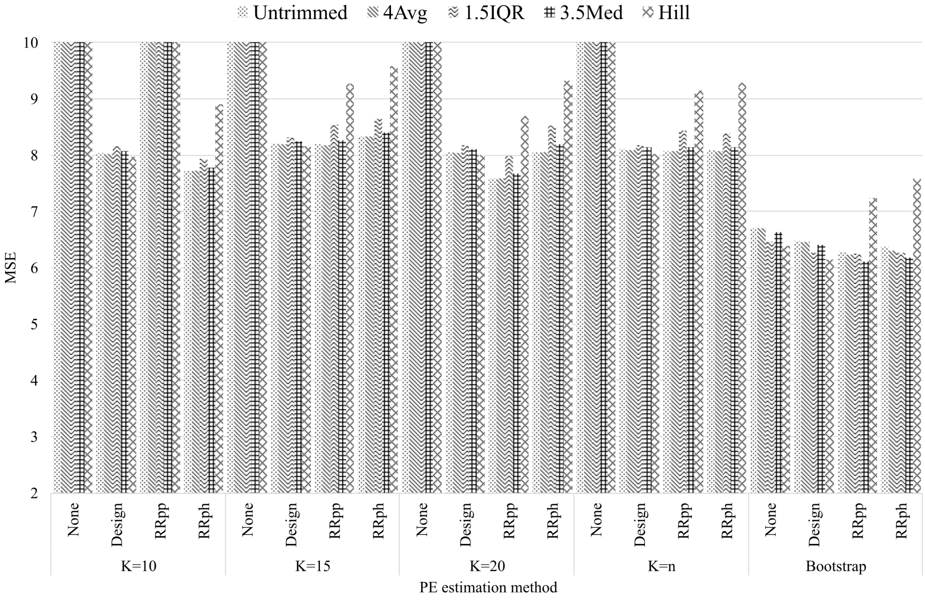

Figure 7. Comparing the “true” MSE of different CV estimators and the bootstrap estimator of PE using the Molinaro approach.

the use of cross-validation and the bootstrap to estimate the PE of the SWLS model under different sampling weights, both trimmed and untrimmed.

To determine which PE estimator performed "best" it was important to determine a "true" PE. Molinaro et al. (2005) proposed selecting a number of samples from a surrogate population and then using these samples as training sets for model fitting. The training models are then used to predict the test set, which consists of all observations in the surrogate population but not in the training set. In this article $R = 100$ samples were selected from the surrogate population (IES 2005/2006) and by the Molinaro approach, 100 "true" PEs were obtained. This is similar to a validation set (VS) approach. In contrast to this the Luus approach was proposed where the average of the 100 "true" PEs was used as a single "true" PE. A "true" PE was obtained for each sampling weight, untrimmed and trimmed, using both approaches. The "true" biases of the PE estimators are the same under both approaches, but overall it was found that the "true" MSEs of the PE estimators are smaller under the Luus "true" PE. Taking the average of the 100 "true" PEs appears to have smoothed the sampling variability resulting in a more stable "true" PE.

Next it was found that at least $K = 15$ splits are required to adequately capture the variance structure of the CS data. It should be mentioned that, since the splitting is done on the PSUs, the number of PSUs would gauge the number of splits you can use. This surrogate population had a large enough number of PSUs that allowed this range of K to be used. The analyst should just ensure that each split contains at least two PSUs. Furthermore, under cross-validation, the Hill trimmed person and household level benchmarked weights resulted in estimated PEs with the smallest bias while the 1.5IQR trimmed person level benchmarked weights resulted in estimated PEs with the smallest MSE. Concerning the bootstrap method, the bias of the bootstrap PE estimator was high in comparison to the bias of the cross-validation PE estimators, but its MSE was lower than the MSE of the cross-validation PE estimators. This implies that the variance of the bootstrap PE estimator is smaller than that of the different cross-validation PE estimators.

In this paper it was 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. As such the test set sampling weights are not adjusted nor are they used in the calculation of the test set PE. 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. Further research would consider this alternative.

With regards to the bootstrap estimator of PE, the bootstrapping pairs approach to linear modelling was used in this paper. Alternatively, the bootstrapping residuals approach can be considered and this would be done under further research.

The simulation results showed that the bootstrap estimator, on average, results in a lower estimated PE (see Figure 1). It appears the bootstrap adjustment to the apparent error is too small in general, leading to a too small final PE estimate. This is an aspect that needs further investigation in order to obtain a more realistic adjustment.

A further observation made from the simulation results is that the bootstrap estimator generally resulted in a larger bias. One suggestion is to use the bootstrap to estimate this bias and then use that result to perform "bias-correction" of the bootstrap PE estimator. Alternatively, Efron and Tibshirani (1998) discuss the .632 bootstrap estimator of PE, but it was not included here due to it being fairly computationally expensive. Given the improvement in modern computing power this estimator might be viable for further research.

Finally, research is also currently underway on the application of the cross-validation methods for the evaluation of the logistic regression model under CS.

Acknowledgement. The authors would like to thank the referee for his/her very thorough reading of the manuscript and for the suggestions that greatly improved the readability of the paper. The suggestion for further research is also greatly appreciated and has been added as a topic for further investigation.

References

Deville, J. and Särndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376–382.

Efron, B. and Tibshirani, R. (1998). *An Introduction to the Bootstrap*. Chapman and Hall, Boca Raton.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York.

Heeringa, S., West, B., and Berglund, P. (2010). *Applied Survey Data Analysis*. Taylor and Francis Group, Boca Raton.

Izrael, D., Battaglia, M., and Frankel, M. (2009). Extreme survey weight adjustment as a component of sample balancing. *SAS Global Forum*, **247**, 1–10.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, New York.

Lohr, S. (2010). *Sampling: Design and Analysis*. Brooks/Cole, Boston.

Luus, R. (2016). *Statistical Inference of the Multiple Regression Analysis of Complex Survey Data*. Ph.D. thesis, Stellenbosch University, Stellenbosch.

Molinaro, A., Simon, R., and Pfeiffer, R. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics*, **21**, 3301–3307.

Neethling, A. and Galpin, J. (2006). Weighting of household survey data: A comparison of various calibration, integrated and cosmetic estimators. *South African Statistical Journal*, 123–150.

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

Rust, K. and Rao, J. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, **5**, 283–310.

Valliant, R., Jever, J., and Keuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York.

Manuscript received 2019-12-09, revised 2020-06-21, accepted 2020-06-24.