Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

SPECIAL ISSUE PAPER

K-fold cross-validation for complex sample surveys

Jerzy Wieczorek* | Cole Guerin | Thomas McMahon

Department of Statistics,
Colby College, Waterville, Maine, USA

Correspondence
*Jerzy Wieczorek, 5841 Mayflower Hill,
Waterville, ME 04901, USA.
Email: jawieczo@colby.edu

Abstract

Although K-fold cross-validation (CV) is widely used for model evaluation and selection, there has been limited understanding of how to perform CV for non-iid data, including from sampling designs with unequal selection probabilities. We introduce CV methodology that is appropriate for design-based inference from complex survey sampling designs. For such data, we claim that we will tend to make better inferences when we choose the folds and compute the test errors in ways that account for the survey design features such as stratification and clustering. Our mathematical arguments are supported with simulations and our methods are illustrated on real survey data.

KEYWORDS:
Survey sampling; Cross validation; Model selection

1 | INTRODUCTION

Predictive modeling with complex sample data is becoming more and more prevalent. Holbrook, Lumley, and Gillen (2020) list 14 different papers building prediction models from a single complex sample survey alone. Consequently, there is a need for valid ways to assess the prediction error of such models. K-fold cross-validation (CV) is one of the most widely applied and applicable tools for model evaluation and selection, but standard K-fold CV relies on an assumption of exchangability which does not hold for many complex sampling designs.

In Section 2, we propose and justify a "Survey CV" methodology that is appropriate for design-based inference from complex survey sampling designs. We claim that we will tend to make better inferences when we choose CV folds and compute test errors in ways that account for the survey design features such as stratification, clustering, and unequal sampling probabilities.

In Section 3, we give two illustrative examples of Survey CV with real data, including a motivating example from our own prior work where complex sample survey data was used to fit a model for predictive use on *future* data not from the survey itself. This is distinct from the traditional goal of model-based estimation for surveys, which is to make inferences about a specific parameter using the survey data at hand. We close with simulations that confirm our intutions about how Survey CV compares to standard approaches when data are not iid.

While survey sampling and CV are both widely used, the overlap in their audiences appears to be small, so readers may be familiar with one but not the other. We summarize K-fold CV in detail in Section 2, but we recommend Section 7.10 of Hastie, Tibshirani, and Friedman (2009) for further details. For a recent review on survey sampling, see the 2017 special issue on complex surveys in *Statistical Science*, including a succinct introduction to complex survey designs and to design-based inference (Skinner & Wakefield 2017).

Briefly, "complex sampling" generally refers to taking a non-iid probability sample from a finite population. In a simple random sample (SRS) without replacement of size n from a population of size N, each set of n observations has the same probability of being selected. Even this is not an iid sample because observations are not independent, as the same unit cannot be drawn twice. More complex sampling designs may feature unequal sampling probabilities, as with "probability proportional to size" (PPS) sampling where each unit in the population is assigned a sampling probability in proportion to some auxiliary variable already known before sampling. Sampling designs may also affect the joint probabilities with which pairs or groups of units are selected. Some such sampling designs tend to increase statistical efficiency, for instance "stratification": partition the population

[0]**Abbreviations:** CV, cross-validation; df, degrees of freedom; iid, independent and identically distributed; MSE, mean squared error; NSFG, National Survey of Family Growth; PPI, Poverty Probability Index; PPS, probability proportional to size; PSU, primary sampling unit; SRS, simple random sample

into subgroups of interest or "strata" and take random samples independently within each stratum, ensuring lower sampling variation than a SRS of the same size. Other sampling designs tend to reduce statistical efficiency but may still be worthwhile because they also lower the cost of data collection, for instance "clustering": partition the population into groups called "clusters" or "Primary Sampling Units" (PSUs), take random samples of these PSUs, and observe all individual units within the selected clusters (or possibly do further subsampling).

For example, strata may be demographic subgroups, ensuring that enough respondents from each subgroup will be represented in the dataset. On the other hand, clusters might be schools and individual units might be schoolchildren; it is less statistically efficient to sample several schools in the USA and observe all children in those schools than to take a SRS of the same number of children, but for a face-to-face survey, the cost of sending interviewers around the entire USA to observe a SRS would be prohibitive. While clustered designs tend to have a smaller "effective sample size" than SRS, it can be possible to achieve a desired standard error more cheaply with a larger cluster sample than with a smaller SRS.

Finally, survey data is often analyzed under a framework of "design-based inference," which assumes that our samples are taken from a finite population in which the data have unknown but constant values. Randomness comes in only through the sampling process, not through a statistical model for how the data are distributed in the full population itself. The bias and variance of an estimator are framed only in terms of how the estimator would vary if we could repeatedly take samples from this population using the same sampling design. We claim that design-based inference is also a useful way to think about CV.

CV is a method for splitting samples into "folds" to directly estimate the out-of-sample performance of predictive models. CV is widely used for choosing tuning parameters or for selecting one model from a set of candidates. Uncorrected in-sample performance would favor the largest or most-flexible models, which tend to overfit to the training sample and generalize poorly to future data. CV corrects for this optimism and avoids overfit by assessing each trained model on held-out data.

For iid data, many other model-selection corrections have been developed, such as AIC (Akaike 1998) or BIC (Schwarz 1978). But CV is popular partly because it does not rely on a likelihood function or stochastic model for the population the data came from, so it works "out of the box" for essentially any predictive-modeling algorithm. For traditional CV to be (approximately) unbiased for the out-of-sample performance, it only requires the assumption that the data are an exchangable sample, such as iid or SRS, from the target population. In this sense, it is natural to think of CV as a design-based estimator under the survey sampling framework—and hence to adapt it for non-SRS designs. This has not been previously addressed for complex sample surveys, although variations of CV such as "block CV" have already been adapted to data with other kinds of dependencies such as temporal, spatial, or hierarchical group structure (Roberts et al. 2017). For general correlation structures (including cluster sampling), Rabinowicz and Rosset (2020) propose using CV with the usual random folds and calculating a debiased estimator of prediction error, but their estimator is limited to linear models.

Statisticians have developed other model-selection tools that account for the sampling design, such as the design-based AIC or dAIC (Lumley & Scott 2015) and the Horvitz-Thompson-Efron estimator (Holbrook et al. 2020). These methods are motivated by parametric models for the data, via the likelihood or a parametric bootstrap, although these methods do not require the set of models under consideration to include the true model. However, our CV-based approach only relies on the sampling design itself and does not require a closed-form likelihood function or parametric model. Holbrook et al. (2020) anticipate an extension of CV prediction error estimation to complex sampling, and Lumley and Scott (2015) do connect dAIC to leave-one-out CV, but not directly to other forms of CV.

Meanwhile, the machine learning literature reflects the need to account for hierarchical structures in data when forming CV folds (Saeb, Lonini, Jayaraman, Mohr, & Kording 2017) and suggests the value of using survey weights to calculate CV performance metrics (Kim 2020). However, we are not aware of prior work framing the choice of folds in terms of survey sampling, nor unifying this idea with the use of survey weights.

In the context of grouped data, where each subject in a study is represented by multiple records (e.g., measured at different times), Saeb et al. (2017) draw a contrast between record-wise and subject-wise CV. In subject-wise CV, a subject's records are all in the same fold together; but under record-wise CV, a subject's records could be in both the training and the test sets. Ignoring the structure of such data leads record-wise CV to underestimate prediction error, because a model that memorizes the subjects at hand might fit the test set well, but it will generalize poorly to future data from new subjects. This is an example of what is known in machine learning and data mining as "leakage" between the training and test sets (Kaufman, Rosset, Perlich, & Stitelman 2012). While our proposed Survey CV method below has a distict rationale from preventing leakage, both are ultimately about making CV more realistic in how it generalizes from the sample to new data.

Note: Kohavi (1995) discussed a different form of stratified CV than in our Survey CV proposal below. In Kohavi's framework, the CV folds are stratified on the response variable after the data have already been collected. Sometimes this may indeed be pragmatically necessary, as when modeling a categorical response variable with some rare categories; without stratifying, some CV folds may have no instances of this response category and the models cannot be trained. Otherwise, however, Kohavi's rationale for stratification appears to be to reduce the variability across repeated runs of CV for a *fixed dataset*. By contrast, our Survey CV framework is meant to reflect the actual variability we would have seen across *different samples* from this population. Hence, we stratify the folds only when the original sample comes from a stratified design.When it does not, we claim that stratified CV will underestimate the true sampling variation and thus also the prediction error.

2 | METHODS

2.1 | What is CV actually estimating?

Working with limited data in a complex world, it is often hopeless to select a "true" model. Even if one exists, it may not be in the set of models under consideration or may be too large to fit with the available data. Instead, we often choose a set of approximate models to compare, and from this set we may wish to seek the model f that would have the lowest loss $L(y, \hat{y})$ on new data from the same population $\mathcal{P}$ after being fitted to our observed sample s. For instance, L may be squared error loss $(y - \hat{y})^2$ for regression problems or logistic loss $y \log(\hat{y}) + (1-y)\log(1-\hat{y})$ for binary classification problems. Thus, we might wish to minimize the risk over new data after fitting f to s to get the trained model $\hat{f}_s$. This risk is also called the "conditional test error" $\text{Err}_s(f|\mathcal{P}) = \mathbb{E}_{(x_{new}, y_{new})\sim\mathcal{P}}[L(y_{new}, \hat{f}_s(x_{new}))]$. But this is hard to estimate without more data or strong assumptions.

Instead, K-fold CV is an estimate of the *mean* of $\text{Err}_s(f)$ across new "similar" training samples s_j^*, sampled using the same design $\mathcal{S}$ and from the same population $\mathcal{P}$ as s (Hastie et al. 2009). We estimate this "expected test error,"

$$Err(f|\mathcal{P}, \mathcal{S}) = \mathbb{E}_{s^*\sim\mathcal{S}}\left[\mathbb{E}_{(x_{new}, y_{new})\sim\mathcal{P}}[L(y_{new}, \hat{f}_{s^*}(x_{new}))]\right] ,$$

by finding the empirical risk on K test sets after fitting f to K training sets:

$$\widehat{Err}_{CV}(f) = \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in test_j} L(y_i, \hat{f}_{train_j}(x_i)) . \tag{1}$$

We often use CV to choose a tuning parameter (such as the lasso penalty λ or a spline's degrees of freedom), then use the selected tuning parameter value to refit the model on the full dataset. Hence we might wish to know $\text{Err}_s(f)$, to decide which tuning parameter value is best for the observed sample s—but we are willing to settle for a good estimate of $\text{Err}(f)$, to decide which tuning parameter value tends to work best for similar samples $s^* \sim \mathcal{S}$. To be useful, $\widehat{\text{Err}}_{CV}(f)$ should not be badly biased for $\text{Err}(f)$. Clearly, $\mathbb{E}\left(\widehat{\text{Err}}_{CV}(f)\right) = \text{Err}(f)$ if (1) the training sets are drawn with the same sampling design $\mathcal{S}$ as s; and if (2) the test set elements (x_i, y_i) are chosen and the sample mean of the L values is calculated in an unbiased way for the population of interest $\mathcal{P}$ where future observations (x_{new}, y_{new}) will be predicted.

SRS CV

Usual CV creates the training sets by randomly partitioning s into K folds and combining $K-1$ of them at a time, then calculates $\widehat{\text{Err}}_{CV}(f)$ as an unweighted mean of all the L values. We will call this "SRS CV." If s really is an exchangeable sample from the target population $\mathcal{P}$, the bias comes only from CV using smaller training sets, whose sample size $n \times (K-1)/K$ is less than s's sample size of n. If K is large, or if the risk is not too sensitive to the training set size for the models under consideration, this bias will be small. In practice, this bias is also often similar across competitive models, with little impact on model selection.

Survey CV

But for complex surveys, forming the training sets and estimating $\widehat{\text{Err}}_{CV}(f)$ should reflect $\mathcal{S}$, the actual sampling design of s. We will call this "Survey CV. " Otherwise, there can be more bias in $\text{Err}(f)$ than just due to a difference in sample sizes. Furthermore, this bias can be quite large and can differ across competitive models, causing poor model selection.

First, if the training sets are not selected similarly to how s was originally sampled, then we would be estimating the expected test error over models $\hat{f}_{train_j}$ fitted to data from a different distribution than the one s came from. Next, if we happen to be comparing two competitive models, where one model includes terms that account for the sampling design (e.g., predictors associated with between-cluster differences) while the other model does not, then the bias in $\widehat{\text{Err}}_{CV}(f)$ can be much higher for the second model—even if both models have similar $\text{Err}(f)$. Finally, if each unit's sampling probabilities are not all equal, the unweighted sample means in $\widehat{\text{Err}}_{CV}(f)$ may be a biased estimate for the true $\text{Err}(f)$.

2.2 | Proposed methodology for "Survey CV": CV with complex sample surveys

2.2.1 | Choice of folds

Our advice on forming Survey CV folds turns out to be equivalent to the advice in Wolter (2007), Chapter 2.4, for forming "random groups" for variance estimation, as well as for the group jackknife in his Chapter 4.5-4.6. Wolter also gives similar advice for the bootstrap in his Chapter 5.3-5.4. Each of these four tools—CV, random groups, jackknife, and bootstrap—estimates some aspect of sampling variability with the help of subsamples that mimic the original sampling design. However, random groups, the jackknife, and the bootstrap are all used to estimate the variance of a statistic from the sample at hand (though the jackknife can also be used for bias reduction). Meanwhile, CV is used to estimate prediction error over future data (although the bootstrap can be used for this purpose as well). Due to its split-sample nature, where test sets are used to

evaluate models fitted to training sets, CV cannot be reduced to a special case of random groups or the jackknife. Likewise, CV (in which the folds are partitions) is not a special case of the bootstrap (where training sets are sampled with replacement).

Nonetheless, Wolter's advice for taking subsamples also applies to forming CV folds:

- For SRS sampling, partition the observations in the dataset completely at random into K equal-sized folds, as usual.
- For cluster sampling, partition the data at the level of the PSUs. All elements from a given PSU should be placed in the same fold, so that the folds are a random partition of PSUs rather than of elementary sampling units. (Note: with multistage sampling, after a first-stage cluster sample there is further subsampling in the selected clusters. Since there is no straightforward way for CV to mimic this subsampling, we simply form folds at the PSU level even in multistage samples.)
- For stratified sampling, make each fold a stratified sample of units from each stratum. Create SRS CV folds separately within each stratum, then combine them across strata.
- For unequal probability samples, such as probability proportional to size (PPS), make each fold a SRS of the data. However, see below for advice on using sampling weights to estimate $\widehat{\mathrm{Err}}_{\mathrm{CV}}(f)$ as a weighted mean and to fit each $\hat{f}_{\mathrm{train}_j}$.
- For more complex sample designs that combine several of the design features above, combine these rules as needed.

The advice for clustered and stratified designs is self-explanatory: each training set should mimic the way the real dataset was sampled, helping CV account for the real data's increased variability due to clustering or reduced variability due to stratification. However, for unequal probability samples, why do we not we take weighted samples when creating folds? We quote Lemma 1 of Cheng, Slud, and Hogue (2010):

Lemma 1. Suppose a sample S is a probability proportional to size (PPS) sample with sample size n drawn from universe U of known size N. Suppose further that the subsample $S_m \subset S$ is to be drawn by simple random sampling taking m out of n. Then, S_m is a PPS sample with size m, and the second-order inclusion probabilities for distinct pairs of elements of the sub-sample are also proportional to the corresponding joint inclusion probabilities for the sample S.

In other words, if we are given a PPS sample and we use SRS to create folds, then each fold will be a subsample with the same PPS sampling design (just at a smaller sample size) as the original sample, as desired.

2.2.2 | Use of weights

For usual CV, $\widehat{\mathrm{Err}}_{\mathrm{CV}}(f)$ is an unweighted sample mean of the n values of the loss, $L(y_i, \hat{f}_{\mathrm{train}_j}(x_i))$. The same n values of L can be used to estimate the standard error of this mean, which is sometimes used in the "1 standard error rule"—instead of selecting the model that minimizes $\widehat{\mathrm{Err}}_{\mathrm{CV}}$, select the sparsest or simplest model whose $\widehat{\mathrm{Err}}_{\mathrm{CV}}$ is within one standard error of the minimum (Breiman, Friedman, Olshen, & Stone 1984).

For Survey CV, we propose using the full dataset's sampling design to replace Equation (1) with a design-based estimate of the mean of the n values of L, as well as to calculate a design-based estimate of this mean's standard error. For instance, the Horvitz-Thompson estimator of the mean can account for sampling design features such as stratification or unequal sampling probabilities (Horvitz & Thompson 1952). As Kim (2020) notes, such design-based estimates of CV error are not intended to show better performance—they are designed to reduce bias in how each model's CV error estimate generalizes to the population, helping us choose the model that will best predict future data.

That said, this advice assumes that the population that was sampled is the same population on which the model's future predictions will be made. If that is not the case, additional care is needed to "transport" the prediction model and assess its expected performance for the target population (Steingrimsson, Gatsonis, & Dahabreh 2021). Sugiyama, Krauledat, and Müller (2007) have adapted CV to "covariate shift," a special case where the distribution of future input points differs from the current dataset, but the conditional distribution of the response given the inputs does not change. However, their work assumes the current dataset was sampled iid, and it has not been adapted yet for Survey CV.

Furthermore, the advice above assumes that the sampling weights are inverse probability weights. In practice, sampling weights are often modified to incorporate adjustments for nonresponse, undercoverage, and post-stratification. More work is needed to determine how Survey CV should incorporate such weighting adjustments.

Finally, the survey sampling design can also be used in the model-fitting process itself within each training set. We acknowledge there is debate about whether and how survey weights should be used in model fitting, depending on the goals of the analysis. See, e.g., Gelman (2007) with discussion, or Section 3.2 of Lumley and Scott (2017). Regardless, after CV the final chosen model is often refit to the full dataset or to new data. In this regard, there should be consistency between how the models are fit during CV training and how the selected model will be refit: If you plan to use (not use) the sampling design to refit the final model, you should also use (not use) it to fit each candidate model during training.

3 | RESULTS

First, we motivate our methodology using illustrative examples on real data. Next, simulations confirm our intuition from Section 2: (1) Typical cluster sampling causes SRS CV to be overconfident, but Cluster CV corrects for this. (2) Typical stratified sampling shows the reverse situation, and Stratified CV corrects for this. (3) With informative sampling weights, using weights in model-training vs. in test-error-estimation can have distinct effects, and both are important.

3.1 | Real data

The anticipated differences between SRS CV and Survey CV can be observed in real data analyses.

First, Figure 1a expands on an intermediate stage in Kshirsagar, Wieczorek, Ramanathan, and Wells (2017)'s work on developing a Poverty Probability Index (PPI) for Zambia, using a logistic-regression group lasso to build a sparse predictive model for household poverty status. The data came from a nationally-representative clustered survey sample of Zambian households from 2015. Kshirsagar et al. used group lasso as part of a process of choosing only a sparse subset of predictor variables that would need to be collected when making predictions for future households. They used CV to choose the lasso tuning parameter, using the sampling weights in fitting models and evaluating test errors. However, at the time, they did not know how CV folds should account for the survey design and simply used weighted SRS CV.

Figure 1a shows how the CV test-set mean logistic loss varies with $-\log(\lambda)$, where λ is the lasso penalty. Smaller $-\log(\lambda)$ values correspond to stricter penalties and sparser models. The CV error estimates are averages (weighted if applicable) across 5 repetitions of 5-fold CV.

Clustered designs like this one are typically less statistically efficient than a SRS, and this is reflected in the CV error curves. Compared to the Unweighted SRS CV error curve, the Weighted Survey CV error curve is shifted up and to the left. In other words, by partitioning entire clusters into folds, Survey CV does better at admitting how hard it is to make good predictions, and avoids overfitting: its CV error minimum is at a smaller $-\log(\lambda)$, leading to a sparser model. The Figure also shows a Weighted SRS CV error curve: sampling weights are used in fitting models on training sets and in calculating CV error on test sets, but folds ignore the survey design. Since this curve is in between the other two, the effect of sampling weights alone is distinct from the effect of accounting for survey design when making CV folds.

Second, Figure 1b makes a similar comparison of CV error curves for a different dataset. We chose the 2015-2017 National Survey of Family Growth (NSFG) for illustrative purposes, since public-use microdata are available and the survey design includes both clusters and strata as well as sampling weights (National Center for Health Statistics 2021). Our subset of the data is representative of women in US households and aged 20-40 years old at time of the survey who have had at least one live birth. In this illustrative example, natural spline models at several degrees of freedom (df) are used to regress the respondent's current income (at time when the survey was taken, measured as a percentage of the poverty threshold) against the respondent's age at first conception.

Figure 1b plots the CV test-set mean square error (MSE) against df. Spline models with more df are more flexible; a 1-df model is a simple linear regression. These CV error estimates are averages (weighted if applicable) across 5 repetitions of 4-fold CV. We could not use 5-fold CV since there were only 4 clusters per stratum in the NSFG survey design.

Although the NSFG does incorporate strata, the use of clusters is likely to make the overall design less statistically efficient than a SRS. Indeed, we see that compared to the Unweighted SRS CV error curve, the Weighted Survey CV error curve is shifted up, although both CV error curves are minimized at the same df value. Likewise, the Weighted SRS CV error curve is roughly between the other two.

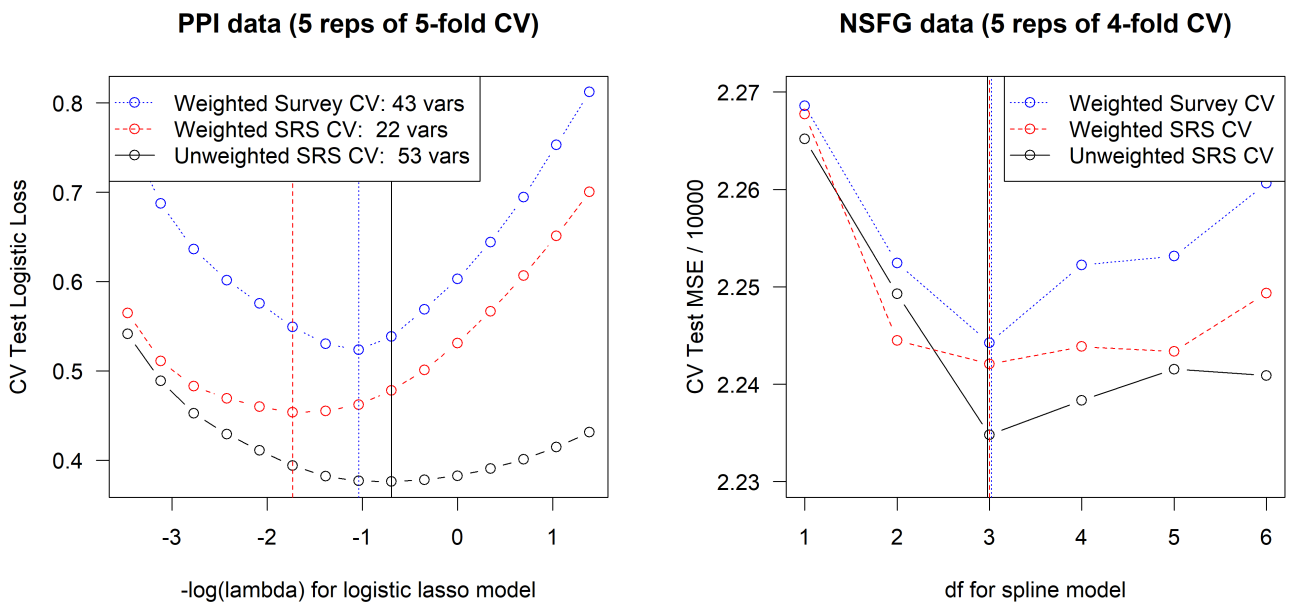
In both subfigures, the Weighted SRS CV error curve tends to have CV errors between the other two curves, but the location of the curve's minimum is erratic: for the PPI it chooses a much smaller model than even Weighted Survey CV does, while for the NSFG it chooses the same model as Unweighted SRS CV does. We expect that Weighted SRS CV is not robust to the possibility of survey-design misspecification interacting with model misspecification. In Section 3.2.2 we further explore the role of sampling weights, using simulated data.

3.2 | Simulations

We simulate a finite population of two continuous variables X and Y, with wide variability around an approximately cubic trend line. We evaluate our proposed CV methodology by taking repeated SRS, clustered, stratified, and PPS samples from this population, and using different CV methods to compare the fits of natural splines with degrees of freedom (df) from 1 to 6. Figure 2 shows the finite artificial population and examples of SRS, clustered, and stratified samples.

3.2.1 | Choice of folds

To simulate informative clustering, we sort the data by X and cluster together consecutive datapoints in groups of 10, giving high homogeneity of both X and Y within clusters. Likewise, to simulate informative stratification, we sort the data by X and partition the population into 10 equal-sized



(a) PPI

(b) NSFG

FIGURE 1 Three types of CV error curves for (a) logistic lasso tuning parameter $-\log(\lambda)$, for model selection in PPI project using 2015 Zambia data; and for (b) spline df, for predicting income from education level using 2015-2017 NSFG data. In both subfigures, lower x-values correspond to smaller / more-penalized models. By using survey folds and weights, Survey CV appropriately accounts for the survey design, showing higher CV error and choosing more-penalized models than SRS CV. Using only survey weights with SRS folds is not adequate.

strata. We then compare taking a SRS of 100 data points, vs cluster sampling by taking a SRS of 10 clusters of 10 points each, vs stratified sampling by taking SRSs of 10 points each from within each of the 10 strata. All observations had equal sampling weights in this set of simulations. For each sampled dataset, we fit splines and report test MSEs using both SRS CV and the appropriate Survey CV. Finally, for each simulated sample, we also trained spline models on an 80% subsample and calculated prediction errors on the full population, then averaged over samples, to assess the true $\mathrm{Err}(\mathrm{f})$ at the 5-fold CV training set size for each df.

Each boxplot in Figure 3 shows a distribution of CV test-set MSEs over 500 such simulations, and each line shows the mean full-population MSE over 500 such simulations. Using SRS CV with SRS samples (top left subfigure) tends to choose 3 df, which is a reasonable choice for the approximately cubic population. The boxplots are slightly above the line, so 5-fold SRS CV slightly overestimates the actual mean prediction error for an SRS of size 80. However, due to within-cluster homogeneity, each cluster sample's effective sample size is much smaller than its nominal $\mathrm{n} = 100$, so we expect that an honest CV method should estimate larger MSEs and choose smaller models for our cluster samples than for SRS data. The reverse should be true for our stratified samples.

For cluster samples, wrongly using SRS CV (top center) gives similar test MSEs as we saw for SRS samples, but now they tend to fall far below the line, showing that SRS CV underestimates the true prediction error for these cluster samples. Correctly using Cluster CV (bottom center) gives higher, more-variable test MSEs, overestimating the true prediction error slightly, and tending to choose smaller splines of only 1 df. Cluster CV is correcting for SRS CV's overconfidence; in fact it appears to be overcorrecting here, although this may be due to the extremely high within-cluster homogeneity. Also, averaging over several repetitions of 5-fold CV would reduce the variability in test MSEs.

Finally, for stratified samples, wrongly using SRS CV (top right) again gives similar test MSEs as for SRS samples, but correctly using Stratified CV (bottom right) gives slightly lower, less-variable test MSEs that are noticeably closer to the line for true prediction errors from these stratified samples. The additional precision gained by stratified sampling appears to allow fitting splines with more than 3 df without overfitting to the training set, and Stratified CV is able to detect this better than SRS CV can.

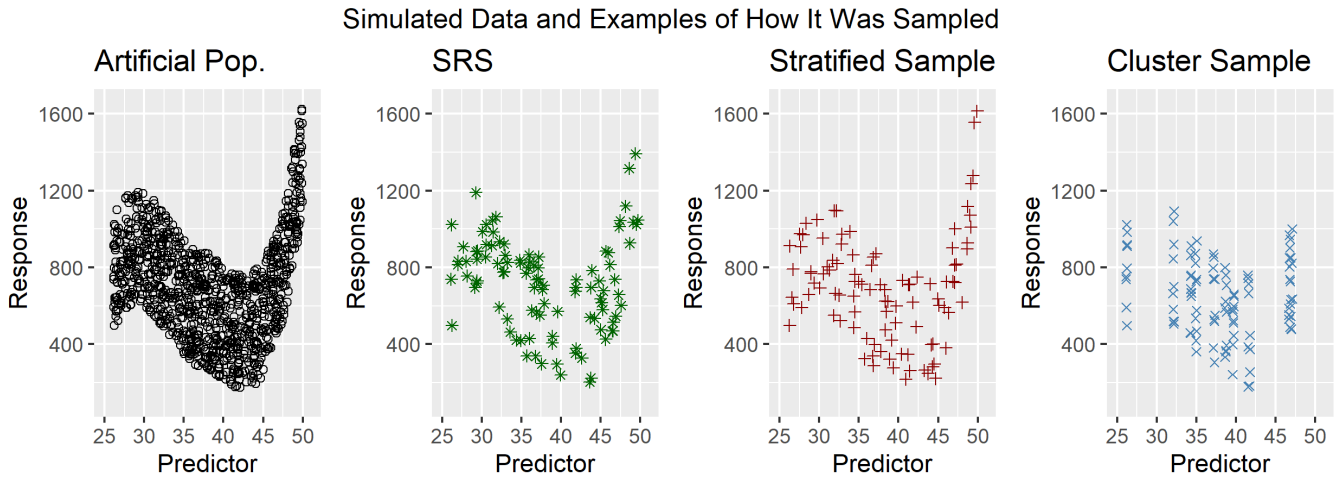


FIGURE 2 Artificial population used in simulations for this paper, and examples of three different types of samples taken from this population.

3.2.2 | Use of weights

To simulate informative survey weighting, we draw a quadratic curve that fits our cubic artificial population poorly, then assign higher sampling probabilities to points nearer this curve, as shown in Figure 4. Next, we take unequal-probability samples of size 100 without replacement, not using any clustering or stratification. We evaluate the fit of natural spline models at several df using four variants of CV. The distributions of CV test-set MSEs from 500 such simulations are shown in Figure 5.

Unweighted SRS CV (bottom subfigure) is naive about the informative sampling and often picks a 2-df model, despite the 3-df population, because the sampling probabilities were higher along a quadratic curve. However, when sampling weights are used in CV both to train models and to calculate test MSEs (top), the MSEs appear more honest: they tend to be higher and more-variable, and the more-appropriate 3 df model tends to win.

In between, we consider two other alternatives. First, sampling weights are used to train models but not to calculate test MSEs (left). For 3 df and above, the weighted training sets lead to fitted models which look more cubic than quadratic—but the unweighted test sets look quadratic and estimate high MSEs for high-df models. This incorrectly causes 2 df to tend to have the lowest test MSEs.

Second, sampling weights are used to calculate test MSEs but not to train models (right). For 2 df and above, the unweighted training sets lead to fitted models which look quadratic—but the weighted test sets are looking for cubic models, leading to a plateau of high test MSEs across the board. This gives the incorrect impression that that there is almost no difference between 2, 3, or more df.

Both the left and right subfigures show, in different ways, that partial use of sampling weights is not enough to correct for the mismatch between the apparently-quadratic sample and the approximately-cubic population. When sampling weights are informative, as they are here, we recommend that the weights be used both for training and for testing.

4 | DISCUSSION

The way that CV folds are chosen can make a real difference in model evaluation and selection when the data come from a complex survey design. We extend the rule-of-thumb in Hastie et al. (2009)—"cross-validation must be applied to the entire sequence of modeling steps"—with our own rule-of-thumb: The sampling design must be taken into account when choosing CV folds and when computing CV test errors. Thanks to these extra precautions, Survey CV allows us to honestly assess which models we can actually afford to fit with the non-iid data at hand.

We expect that the Survey CV framework could be extended to other non-iid sampling methods, such as respondent-driven sampling (RDS), a form of nonprobability sampling where initial respondents are asked to recruit new participants. RDS can be useful for studying hard-to-reach populations, and Gile, Beaudry, Handcock, and Ott (2018) note that "Methods for multivariate (regression) modeling and testing are underdeveloped and in great demand" for RDS.

We have prepared a draft R package, `surveyCV`, to make our methodology widely available to users of the `survey` package (Lumley 2020). Our package currently handles linear and logistic regression models with single-stage clustered or stratified designs, and we will continue to add more

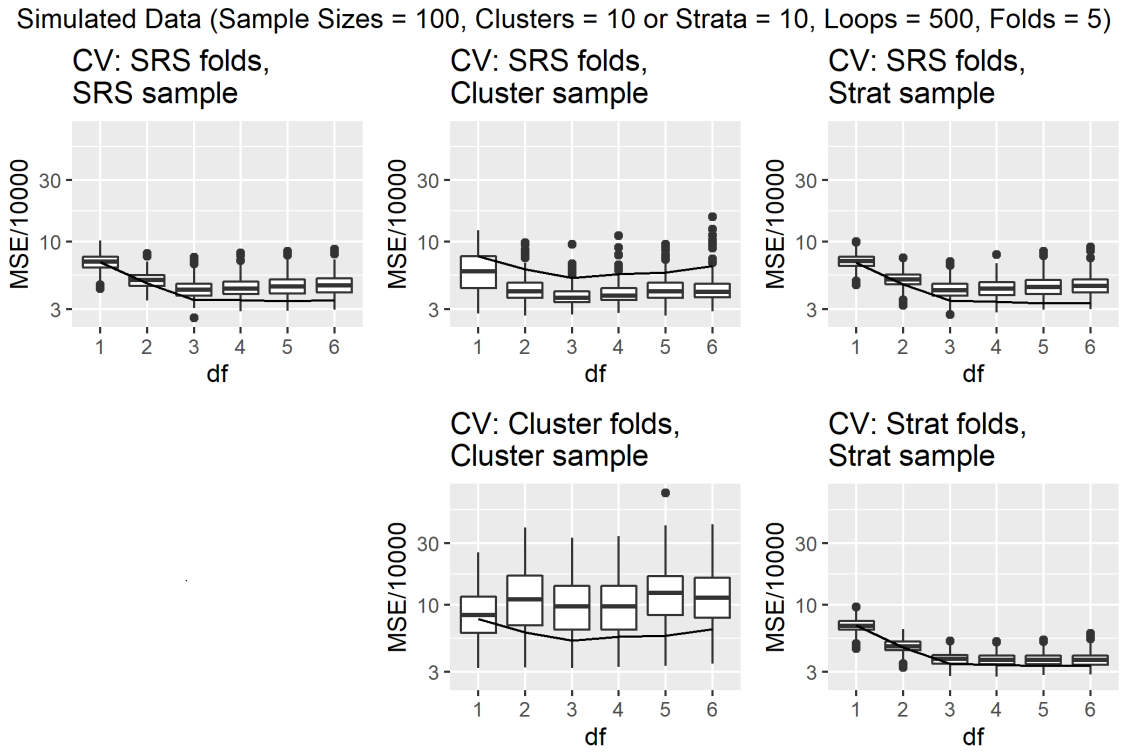


FIGURE 3 Top row shows SRS CV error estimates, bottom row shows Survey CV error estimates; columns show samples each from SRS, clustered, or stratified designs. Each boxplot shows CV error estimates across 500 simulated samples of size 100. Lines show full-population prediction errors averaged over 500 training samples of size 80, the same size as the 5-fold CV training sets. All y-axes are drawn on log scale.

complex designs and models. We also intend to work on ways that Survey CV can account for adjustments to the sampling weights, such as post-stratification and nonresponse adjustments. We will also pursue a better understanding of when we can expect to see a large difference between usual CV and Survey CV, including how it relates to the interplay between model misspecification and design misspecification.

Data availability statement

Our datasets and simulation code are available in our draft `surveyCV` R package on GitHub at `https://github.com/ColbyStatSvyRsch/surveyCV` except for the proprietary 2015 Zambia dataset. While we cannot share that dataset, our R package includes the code used to create Figure 1a.

Acknowledgements

We thank colleagues for their feedback on earlier drafts of this work, particularly Thomas Lumley, Jerry Maples, William Bell, and Michael Collins. Research reported in this publication was supported by a grant from the McVey Data Science Initiative at Colby College.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*. Springer.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Cole Statistics/Probability Series.

Cheng, Y., Slud, E., & Hogue, C. (2010). *Variance estimation for decision-based estimators with application to the Annual Survey of Public Employment and Payroll. Governments Division Report Series, Research Report #2010-3* (Tech. Rep.). U.S. Census Bureau. Retrieved from `https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.225.6683&rep=rep1&type=pdf`

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153–164.

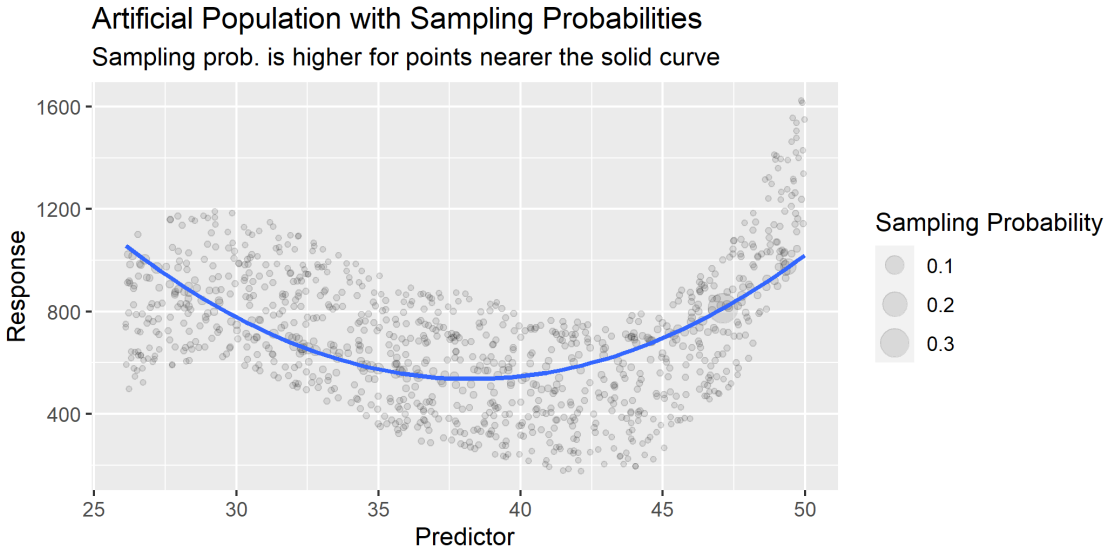


FIGURE 4 Artificial population, with unequal sampling probabilities inversely proportional to vertical distance from the quadratic line shown.

Gile, K. J., Beaudry, I. S., Handcock, M. S., & Ott, M. Q. (2018). Methods for inference from respondent-driven sampling data. *Annual Review of Statistics and Its Application*, *5*, 65–93.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer Science & Business Media.

Holbrook, A., Lumley, T., & Gillen, D. (2020). Estimating prediction error for complex samples. *Canadian Journal of Statistics*, *48*(2), 204–221.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685.

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, *6*(4), 1–21.

Kim, B. (2020). Machine learning model selection with complex sample survey data. In *2020 Symposium on Data Science and Statistics*.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI* (Vol. 14, pp. 1137–1145).

Kshirsagar, V., Wieczorek, J., Ramanathan, S., & Wells, R. (2017). Household poverty classification in data-scarce environments: a machine learning approach. In *NeurIPS 2017 Workshop on Machine Learning for the Developing World*. Retrieved from `arXiv:1711.06813`

Lumley, T. (2020). *survey: analysis of complex survey samples*. R package version 4.0.

Lumley, T., & Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, *3*(1), 1–18.

Lumley, T., & Scott, A. (2017). Fitting regression models to survey data. *Statistical Science*, 265–278.

National Center for Health Statistics. (2021). *National Survey of Family Growth, 2015-2017. Public-use data files and documentation*. Retrieved from `https://www.cdc.gov/nchs/nsfg/nsfg_2015_2017_puf.htm`

Rabinowicz, A., & Rosset, S. (2020). Cross-validation for correlated data. *Journal of the American Statistical Association*. doi: 10.1080/01621459.2020.1801451

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929.

Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). The need to approximate the use-case in clinical machine learning. *Gigascience*, *6*(5), 1–9.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.

Skinner, C., & Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science*, *32*(2), 165–175.

Steingrimsson, J. A., Gatsonis, C., & Dahabreh, I. J. (2021). Transporting a prediction model for use in a new target population. *arXiv preprint arXiv:2101.11182*.

Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, *8*(5).

Wolter, K. (2007). *Introduction to Variance Estimation* (2nd ed.). Springer Science & Business Media.

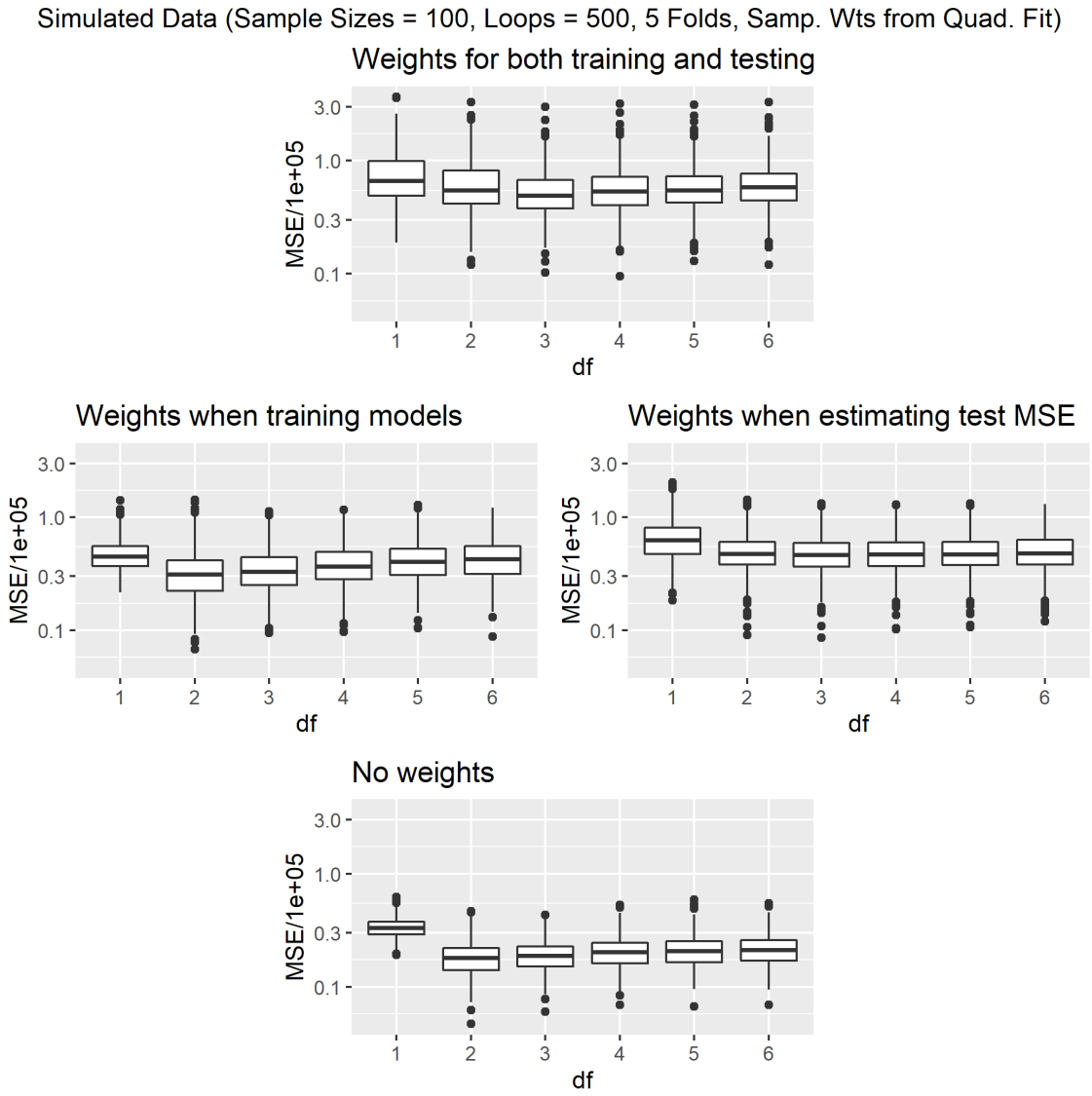


FIGURE 5 CV error estimates, accounting for vs ignoring unequal sampling weights (in model fits and in test MSEs). Each boxplot shows CV error estimates across 500 simulated samples of size 100. All y-axes are drawn on log scale.

How to cite this article: Wieczorek J, Guerin C, McMahon T. K-fold cross-validation for complex sample surveys. *Stat*. *2021;00:1–10*.