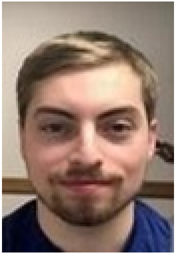
K-Fold Cross-Validation for Complex Sample Surveys

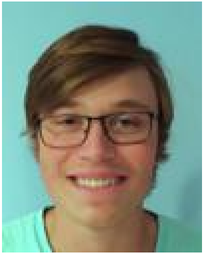
Jerzy Wieczorek
Colby College
jerzy.wieczorek@colby.edu

June 4, 2021
Symposium on Data Science & Statistics

Joint work with Colby College undergraduates



Cole Guerin '21



Thomas McMahon '21

Motivating example: Poverty Probability Index

To decisively know a household's poverty status, need long assessments & trained interviewers: high costs, response burden

ppi
POVERTY PROBABILITY INDEX

A "poverty measurement tool" for organizations serving the poor: Quick & simple country-specific models estimate prob. that a household is below local poverty line

Developing vs. using PPI for a given country

The PPI central office will:

- Obtain recent, nationally-representative household survey data from a nation's statistical agency
- Fit a (penalized logistic regression) model, using a small subset of survey Qs to predict household poverty status
  (see Kshirsagar, Wieczorek, et al. (2017))

Then PPI's "clients" can:

- Carry out own surveys among the communities they serve
- Apply PPI's model to that data to predict poverty status: can target interventions or track overall poverty rates

Example scorecard

Higher total: higher prob. of being above poverty line

Indicator	Value	Points	Score
1. How many household members are aged 25 or younger?	A. 3 or more	0	8
	B. 0, 1, or 2	8	
2. How many household members aged 6 to 17 are currently attending school?	A. Not all	0	0
	B. All	8	
	C. No children aged 6 to 17	21	
3. What is the material of the walls of the house?	A. Mud/cow dung; grass/sticks/makuti; or no data	0	5
	B. Other	5	
4. What kind of toilet facility does your household use?	A. Other	0	2
	B. Flush to sewer; flush to septic tank; pan/bucket; covered pit latrine; or ventilation improved pit latrine	2	
5. Does the household own a TV?	A. No	0	0
	B. Yes	16	

Choosing survey Qs and tuning parameters

The (survey-weighted, elastic-net logistic regression) model has tuning parameters, usually chosen by **cross-validation**.

But:

- Cross-validation usually treats the data as iid, and splits into folds at random before training and testing the models.
- PPI datasets come from complex survey designs, where observations were **not** sampled independently.

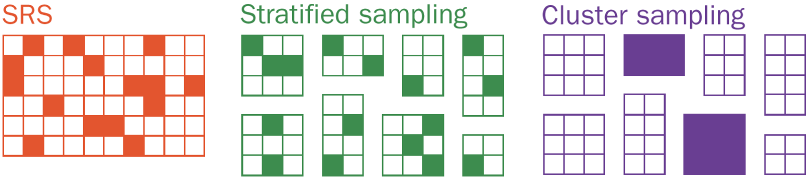
Does this matter???

What are complex survey designs?

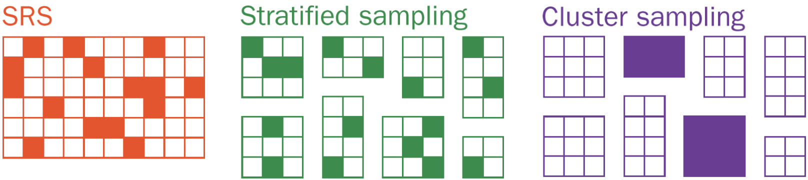
SRS: simple random sampling

Stratified sampling: partition population into "strata," and take samples separately within each stratum

Cluster sampling: partition population into "clusters," and take a sample of clusters, observing all units in each sampled cluster



Complex survey designs: PPI example



National surveys often use:

- sub-national regions as strata—ensures each region gets sampled, and improves statistical precision
- towns or villages as clusters (within strata)—lowers interviewer travel costs, but also reduces precision

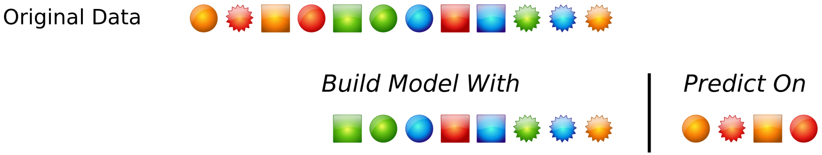
Review: what is data splitting?

Check model predictions on held-out testing data,
to avoid overfitting to the training data.

Partition the data at random into a training set *train*,
used to fit models $\hat{f}_{train}$,
and a testing set *test*, used to evaluate the trained model:

$$\widehat{MSE}(f) = \frac{1}{n_{test}} \sum_{i \in test} \left(y_i - \hat{f}_{train}(x_i)\right)^2$$

Pick a model f with low $\widehat{MSE}(f)$, or other expected loss $L(y, \hat{y})$.



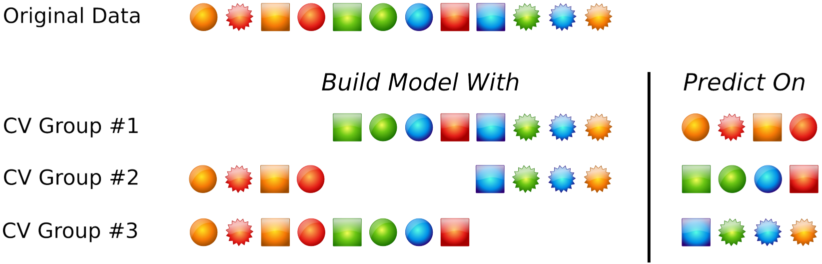
{Image source: Kuhn and Johnson (2013), *Applied Predictive Modeling*}

Review: what is K-fold CV?

Partition the data at random into K equal-sized "folds."
Each training set $train_j$ is the union of $K-1$ folds, and
each held-out fold $test_j$ is used for testing the trained model $\hat{f}_{train_j}$:

$$\widehat{MSE}_j(f) = \frac{1}{n_{test_j}} \sum_{i \in test_j} \left(y_i - \hat{f}_{train_j}(x_i)\right)^2$$

$$\widehat{MSE}_{CV}(f) = \frac{1}{K} \sum_{j=1}^{K} \widehat{MSE}_j(f)$$



{Image source: Kuhn and Johnson (2013), *Applied Predictive Modeling*}

What is CV actually doing?

Possible goals when using CV for model selection:

- Goal 1: Choose the “true” model that best matches the **population**

What is CV actually doing?

Possible goals when using CV for model selection:

- Goal 1: Choose the “true” model that best matches the **population**
  - *(We may not have enough data to do this)*

What is CV actually doing?

Possible goals when using CV for model selection:

- Goal 1: Choose the "true" model that best matches the **population**
  - *(We may not have enough data to do this)*
- Goal 2: Choose the best model we can *afford* to fit with **this specific sample**

What is CV actually doing?

Possible goals when using CV for model selection:

- Goal 1: Choose the "true" model that best matches the **population**
  - *(We may not have enough data to do this)*
- Goal 2: Choose the best model we can *afford* to fit with **this specific sample**
  - *(Hard to do without strong assumptions or extra data)*

What is CV actually doing?

Possible goals when using CV for model selection:

- Goal 1: Choose the "true" model that best matches the **population**
  - *(We may not have enough data to do this)*
- Goal 2: Choose the best model we can *afford* to fit with **this specific sample**
  - *(Hard to do without strong assumptions or extra data)*
- Goal 3: Choose the best model we can afford to fit on samples **like this one**

What is CV actually doing?

Possible goals when using CV for model selection:

- Goal 1: Choose the “true” model that best matches the **population**
  - *(We may not have enough data to do this)*
- Goal 2: Choose the best model we can *afford* to fit with **this specific sample**
  - *(Hard to do without strong assumptions or extra data)*
- Goal 3: Choose the best model we can afford to fit on samples **like this one**
  - *(This is what CV actually approximates)*

    (see Hastie et al., *Elements of Statistical Learning*, Ch 7)

What is CV actually doing?

Instead of risk (expected loss $L(y, \hat{y})$) for the observed sample s,

$$Err_s(f) = \mathbb{E}_{(x_{new}, y_{new})} L(y_{new},\ \hat{f}_s(x_{new}))\,,$$

K-fold CV tries to estimate average risk over similar samples s^*

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

as empirical risk on K test sets after fitting f to K training sets:

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

The way CV selects train/test sets affects bias of $\widehat{Err}_{CV}(f)$.
For usual CV, bias is only from training set sizes: $n \times \frac{K-1}{K} < n$.

Why not use usual CV for complex survey designs?

- If s was iid sample of size n, usual CV's bias in $\widehat{Err}_{CV}(f)$ only comes from training set size $n \times \frac{K-1}{K} < n$. Often this bias is (a) small and (b) nearly constant across competitive models, so it should not affect model selection much.

Why not use usual CV for complex survey designs?

- If s was iid sample of size n, usual CV's bias in $\widehat{Err}_{CV}(f)$ only comes from training set size $n \times \frac{K-1}{K} < n$. Often this bias is (a) small and (b) nearly constant across competitive models, so it should not affect model selection much.
- But for complex surveys, each $train_j$ should be formed in a way that reflects **actual sampling design** of s. Otherwise, the bias in $\widehat{Err}_{CV}(f)$ could be (a) large and (b) very different across competitive models, causing poor model selection.

Why not use usual CV for complex survey designs?

- If s was iid sample of size n, usual CV's bias in $\widehat{Err}_{CV}(f)$ only comes from training set size $n \times \frac{K-1}{K} < n$. Often this bias is (a) small and (b) nearly constant across competitive models, so it should not affect model selection much.
- But for complex surveys, each $train_j$ should be formed in a way that reflects **actual sampling design** of s. Otherwise, the bias in $\widehat{Err}_{CV}(f)$ could be (a) large and (b) very different across competitive models, causing poor model selection.
- For complex surveys, when survey respondents don't all have the same sampling probability, bias can also come from taking a simple mean of the loss over test cases.

How **should** we do CV with complex survey data?

1. Create complex-survey CV folds in the same way that we form "Random Groups" for variance estimation & for group jackknife (see Wolter, *Introduction to variance estimation*, Section 2.4)

- For single-stage SRS, divide the sample at random into K folds (as in usual CV).
- For cluster sampling, sample the clusters as units: all elements from a given cluster should be placed in the same fold.
- For stratified sampling, make each fold a stratified sample of units from each stratum.
- For multi-stage sampling, combine these rules as necessary.

How **should** we do CV with complex survey data?

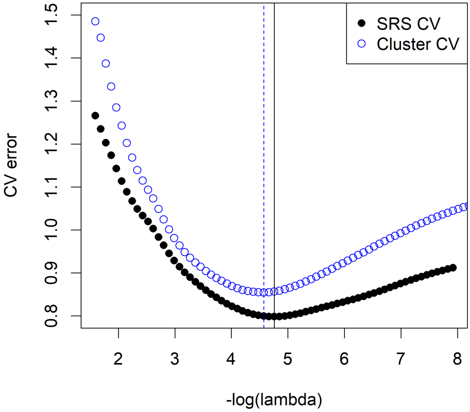
1. Create complex-survey CV folds in the same way that we form "Random Groups" for variance estimation & for group jackknife
   (see Wolter, *Introduction to variance estimation*, Section 2.4)

- For single-stage SRS, divide the sample at random into K folds (as in usual CV).
- For cluster sampling, sample the clusters as units: all elements from a given cluster should be placed in the same fold.
- For stratified sampling, make each fold a stratified sample of units from each stratum.
- For multi-stage sampling, combine these rules as necessary.

2. Account for strata, clusters, survey weights, etc. in calculating expected loss, e.g. use survey-weighted mean for $\widehat{MSE}$.

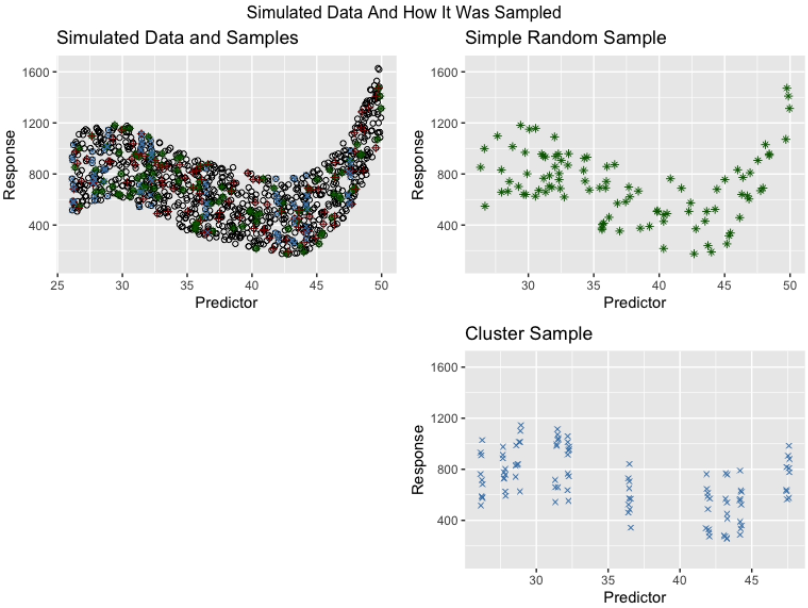
Does it really make a difference? PPI example

Using CV to choose tuning parameter λ in logistic-regression lasso, for PPI for Zambia using a 2015 cluster sample:



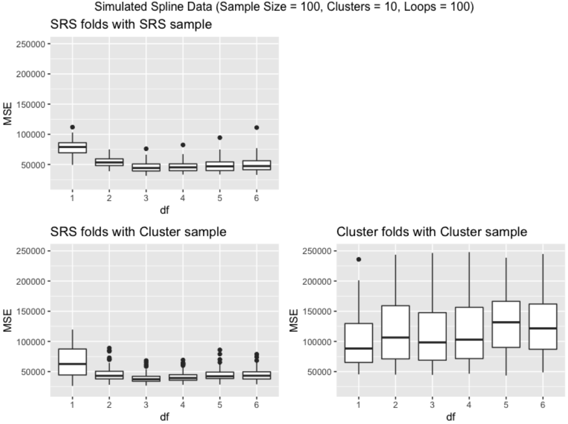
Cluster CV sensibly estimates higher errors and is minimized at a smaller $-\log(\lambda)$ (smaller model) than SRS CV.

Sims: population, and SRS or cluster sampling



Sims: when folds do/don't account for clustering

{Take a sample. Use 5-fold CV to estimate MSEs for splines with df from 1 to 6.} Repeat many times.



On cluster samples, Cluster CV sensibly estimates higher errors and is minimized at a smaller df (smaller model) than SRS CV.

A heuristic for cross-validation

1. **Forming folds: training data should mimic the real sampling design as well as possible, just with smaller n**
   - *Keep same strata and cluster structure – just fewer samples per stratum and fewer of the clusters*
2. **Estimating loss: generalize from testing data to the full population as well as possible**
   - *Use strata, clusters, and weights to compute $\widehat{MSE}$ etc.*

A heuristic for cross-validation

1. **Forming folds: training data should mimic the real sampling design as well as possible, just with smaller n**
2. **Estimating loss: generalize from testing data to the full population as well as possible**

Examples:

- Snowball sampling: all contacts resulting from an initial respondent should be in the same fold
- Panel study: all time points for a respondent should be in the same fold (Saeb et al., 2017)

Conclusion

If data came from a complex survey design, we should account for this when creating cross-validation folds. We will avoid overconfidence and more realistically evaluate how well our model is likely to work when trained on the available data.

To do:

- Better understand complex-survey CV's properties
- More clearly demonstrate its impact on real datasets
- Publish `surveyCV` R package, which extends Thomas Lumley's `survey` package
- Compare with alternatives, such as iid folds but debiased $\widehat{MSE}$ (Rabinowicz and Rosset, 2020) instead of debiased folds

Thank you!

Please reach out, especially if you know of...

- previous literature I've missed on this topic
- datasets that could be a good test case for `surveyCV`
- other study designs on which to try out this CV heuristic

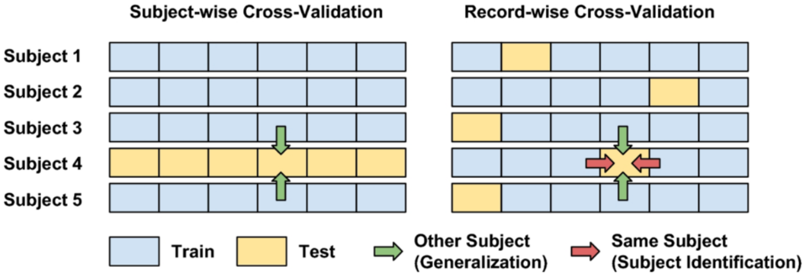
Contact: jerzy.wieczorek@colby.edu or @civilstat

Related work:

- Creel, D. (2019), "Statistical learning for complex survey data: using cross-validation for variable selection in generalized linear models," *GASP*.
- Holbrook, A., T. Lumley, and D. Gillen (2020), "Estimating prediction error for complex samples," *CJS*.
- Kim, B. (2020), "Machine learning model selection with complex sample survey data," *SDSS*.
- Lumley, T. and A. Scott (2015), "AIC and BIC for modeling with complex survey data," *JSSM*.
- Rabinowicz, A. and S. Rosset (2020), "Cross-validation for correlated data," *JASA*.
- Saeb, S. et al. (2017), "The need to approximate the use-case in clinical machine learning," *GigaScience*.

Supplemental slides

Subject-wise vs Record-wise CV



{Saeb et al., 2017}

How is this different than existing stratified CV?

Since at least Kohavi (1995), "stratified CV" for classification problems has been used to mean:
Creating folds by stratifying on the response variable.

This ensures that folds have "balanced classes" – every training and test set has the same distribution of response classes as the full dataset. The heuristic rationale seems to be:

- Every fold should look like the full dataset (but smaller). This will **reduce variability over partitions**, for a **given dataset**.

But this is different from the heuristic that I recommend:

- Every fold should mimic a new (but smaller) sample from the same population, using the sample sampling design. This will more **honestly reflect variability** across **new datasets** we could have gotten, telling us how big a model we can realistically afford to fit.

What about sampling weights?

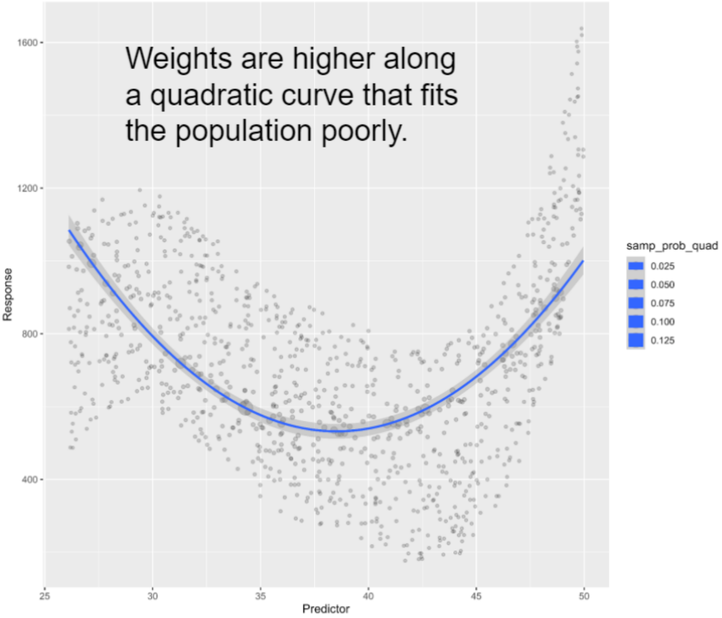
If we know sampling probabilities (or otherwise have survey weights), use them in estimating empirical risk. Recall:

$$\widehat{Err}_{CV}(f) = \frac{1}{K}\sum_{j=1}^{K} \hat{\mathbb{E}}_{(x_{test_j}, y_{test_j})} L(y_{test_j},\ \hat{f}_{train_j}(x_{test_j}))$$

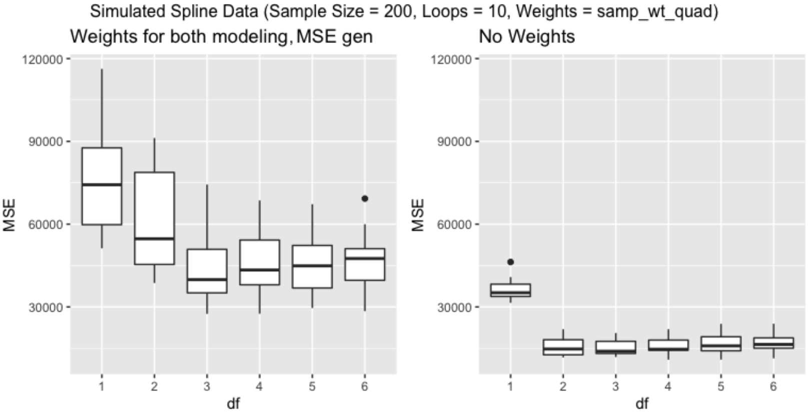
Then $\hat{\mathbb{E}}_{(x_{test_j}, y_{test_j})} L(\ldots)$ can be computed as a survey estimate of a "population mean" of L, generalizing from this sample test set to the population it came from. Use Horvitz-Thompson (inverse probability weighted mean of L across the test set) or other appropriate estimate of population mean.

Most likely, also should use sampling design / weights to fit $\hat{f}_{train_j}$, but that's a separate issue.

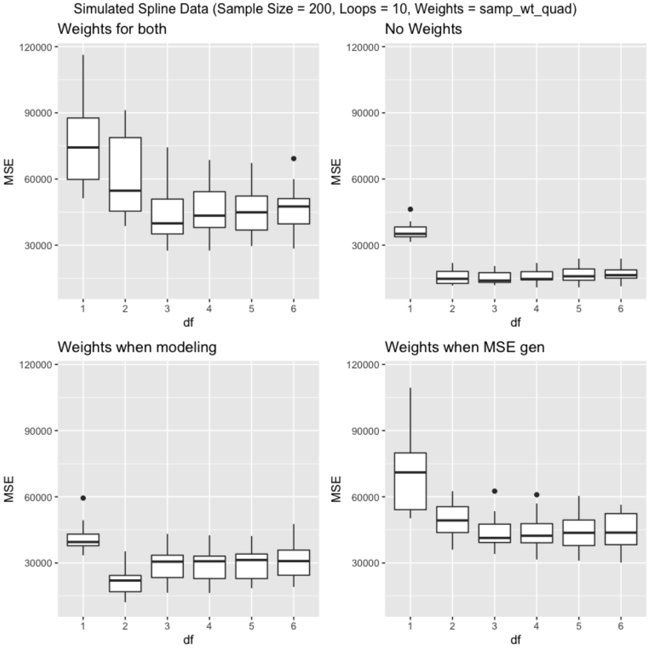
Extra sims: population and weighted sampling



Sims: when MSEs do/don't account for sampling weights



Sims: in further detail



Extra example: NSFG

Using a subset of the 2015-2017 National Survey of Family Growth data, as cleaned by Hunter Ratliff. The survey design has both clustering and stratification.

Fit splines with df from 1 to 6 to predict Income (as % of poverty level) from Years of Education.