2018

European Conference on Quality in Official Statistics

Exploiting auxiliary data: Random Forest Regression estimator

Luis Sanguiao, INE (Spanish NSI), luis.sanguiao.sande@ine.es

June 29th

Session 29

100 lat GUS Statistics Poland

eurostat

Stating the problem

- In addition to the survey sample, we have some additional data from an external source (administrative data, Big Data, …)
- These data, *A* are linked to our population at statistical unit level
- We want to improve the estimations of our target variables exploiting *A*. The idea is $E(X|A)$ should have less variance than $E(X)$.
- That way, we can reduce the sample size and therefore the response burden.
- A machine learning approach is preferred, both to save the analyst's time and to avoid subjectivity.

The Random Forest Regression estimator

- Definition for simple stratified sampling:

$$\hat{X}_{RFRE} = \sum_{i=1}^{N} \hat{x}_i + \sum_{h=1}^{L} \sum_{i=1}^{n_h} N_h \frac{x_{hi} - \hat{x}_{hi}}{n_h}$$

- The out-of-bag prediction is used for in sample units.
- The formula also works for any bootstrap aggregated algorithm. We also know general unbiased model based estimators both for totals and its variance.
- Random Forest and similar algorithms are a good general purpose choice, since they are non linear and non parametric.

Properties

- The RFR estimator is approximately unbiased: it is an approximation of an unbiased estimator.
- An approximate estimator of its variance is

$$\hat{V}_{RFR} = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{e,h}^2}{n_h}$$

- This second approximation is not as good as the first one, so the bias is small but perceptible.
- We know an unbiased version of the estimators, but we currently do not have a fast enough implementation to run the simulations.

Synthetic data simulation (I)

Linear model

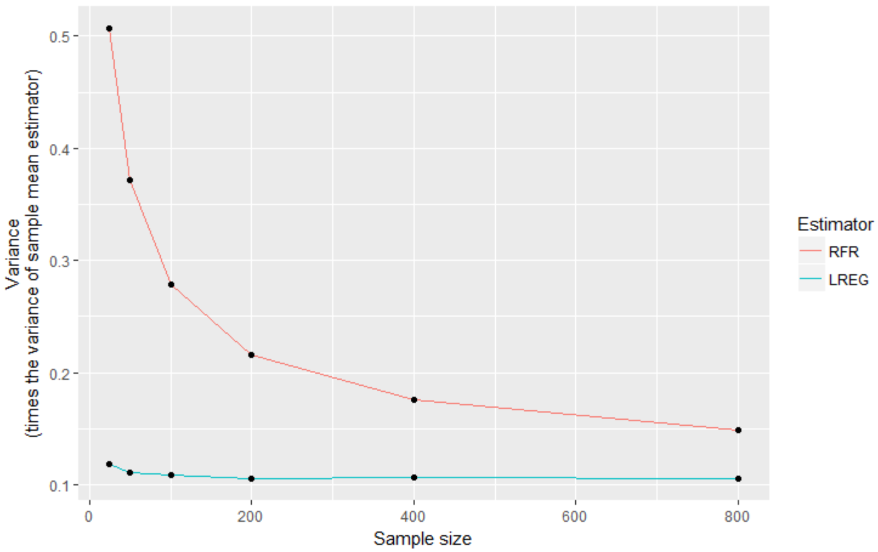
Estimator	RFRE	Linear Regression	Sample mean
Bias (estimator)	0.01	0.11	-0.07
Variance (estimator)	27.88	10.87	100
Bias (estimated variance)	-2.92	-3.33	0.59
Variance (estimated variance)	10.97	0.07	100

Linear model on logarithms

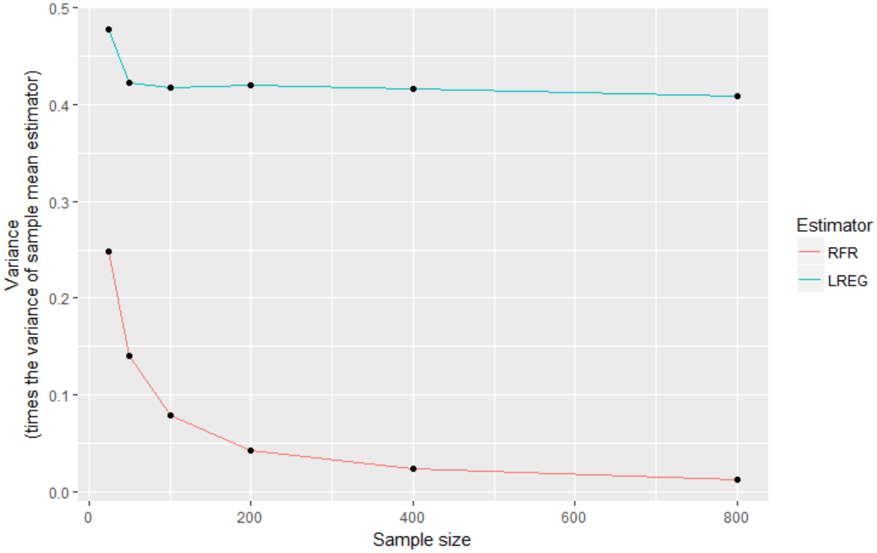
Estimator	RFRE	Linear Regression	Sample mean
Bias (estimator)	0.10	1292.15	-20.14
Variance (estimator)	7.86	41.70	100
Bias (estimated variance)	-4.86	-20.70	-0.20
Variance (estimated variance)	0.98	44.37	100

Synthetic data simulation (II)

Linear model



Linear model on logarithms



- The linear regression estimator learns faster and gets stuck.
- While the RFR estimator continues learning, its variance decreases faster than $O\left(\frac{1}{n}\right)$.

100 lat GUS Statistics Poland   2018 European Conference on Quality in Official Statistics   eurostat

Simulation with real data (SBS)

Total Expenses, simple random sampling

Estimator	RFRE	Linear Regression	Sample mean
Bias (estimator)	0.24	2.77	0.32
Variance (estimator)	60.98	78.73	100
Bias (estimated variance)	-0.97	-24.02	1.83
Variance (estimated variance)	82.82	87.76	100

Total Personnel Expenses, stratified sampling, MV allocation

Estimator	RFRE str	Sample mean str	RFRE	Sample mean
Bias (estimator)	-0.08	0.01	-0.01	-0.04
Variance (estimator)	40.65	86.01	13.60	100
Bias (estimated variance)	-0.89	0.56	-2.05	-0.57
Variance (estimated variance)	2875.59	12300.93	29.88	100

Conclusions

- A good choice when the analyst has a lot of auxiliary data and little idea about how to use it (or little time).
- If the sample is big enough its performance will be similar to those of the *real model*.
- It might be advisable to use the unbiased formula for the variance. The bias is small but still noticeable.
- To preserve affine identities a multivalued version of the Random Forest can be used.
- Up to a certain degree it makes unnecessary the use of a complicated sampling design to improve accuracy.
- Possible additional uses: imputation, small areas, …

Bibliography

- Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5-32.

- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*, Springer, New York.

- Sanguiao (2018). A design unbiased model assisted estimator for finite populations. (Unpublished)

Exploiting auxiliary data: Random Forest Regression estimator

Luis Sanguiao, INE (Spanish NSI), luis.sanguiao.sande@ine.es