*Journal of Survey Statistics and Methodology (2024) **12**, 305–310*

DISCUSSION OF THE 2022 HANSEN LECTURE: "THE EVOLUTION OF THE USE OF MODELS IN SURVEY SAMPLING"

F. JAY BREIDT*

The 2022 Hansen Lecture gave a broad overview of the use of models in survey sampling, with emphasis on modeling approaches to incorporating auxiliary information in survey estimators. This discussion expands upon some issues in model-assisted estimation, exploring data needs and the availability of multipurpose weights for advanced modeling methods.

KEY WORDS: Auxiliary information; Model-assisted estimation; Multipurpose weights.

Statement of Significance

Model-assisted estimation is a general class of methods for incorporating population auxiliary information into survey estimators. Flexible models and methods employed in such estimators include linear models, linear mixed models, kernel regression, splines, additive models, neural nets, shrinkage and selection procedures, and tree-based methods, among many others. Even with these advanced modeling methods, the resulting model-assisted estimators can typically be expressed (at least approximately) in terms of weighted estimates with multipurpose weights, applicable to any study variable.

It is my honor to have the opportunity to discuss this article, which formed the basis of the 2022 Hansen Lecture by Professor Valliant. While the article does an excellent job of summarizing the technical content reviewed in the lecture, it does not fully capture Professor Valliant's rich anecdotes and personal remi-

F. Jay Breidt is a Senior Fellow in the Department of Statistics and Data Science, NORC at the University of Chicago, 55 East Monroe Street, Chicago, IL 60603, USA.

*Address correspondence to F. Jay Breidt, Department of Statistics and Data Science, NORC at the University of Chicago, 55 East Monroe Street, Chicago, IL 60603,USA; E-mail: breidt-jay@norc.org.

https://doi.org/10.1093/jssam/smad030 Advance access publication 26 July 2023

niscences about Morris Hansen that made the live presentation such a pleasure to attend. Special thanks to the organizers and supporters of the Hansen Lecture series for this great service to those of us interested in the theory and practice of surveys.

In my experience, this survey interest group is a pragmatic bunch, always willing to derive benefits from models whenever it makes sense to do so. Models are extremely useful for organizing and communicating thoughts, for deriving estimators with good properties, for assessing expected behavior under ideal conditions, and for identifying nonideal conditions. At the same time, the survey community maintains healthy skepticism about those models. In a production environment for a complex survey, methods have to work again and again: for case after case (often thousands) and study variable after study variable (often hundreds). Estimation techniques need to be robust under model misspecification because methods are often applied generically (e.g., the same calibrated survey weights applied to all study variables) and any proposed model is certainly misspecified for some study variables.

My own research and practice has tended to emphasize model-assisted (MA) estimation: given covariates $\boldsymbol{x}_k$, specify a working model $y_k = \mu(\boldsymbol{x}_k) + \varepsilon_k$, with $\{\varepsilon_k\}$ independent and identically distributed $(0, \sigma^2)$; write down the (infeasible) "estimator" $m_N(\cdot)$ of $\mu(\cdot)$ that would be computed if the entire population were observed; create a feasible plug-in survey-weighted estimator $\widehat{m}_N(\cdot)$; and construct an MA estimator of the y-total as a model-based prediction plus a design-bias adjustment:

$$\mathrm{MA}(y_k) = \sum_{k \in U} \widehat{m}_N(\boldsymbol{x}_k) + \sum_{k \in s} \frac{y_k - \widehat{m}_N(\boldsymbol{x}_k)}{\pi_k}.$$

Under mild conditions, such MA estimators tend to be asymptotically design-unbiased and consistent even if the model is misspecified, with smaller variance than the Horvitz–Thompson estimator $\mathrm{HT}(y_k) = \sum_{k \in s} y_k \pi_k^{-1}$ if the model is reasonably specified. Many MA estimators can be constructed with this basic recipe; see Breidt and Opsomer (2017) for a partial review.

One other potential advantage of the MA approach is relevant in cases where inclusion probabilities π_k are not completely known but require some model specification and estimation. This can occur due to coverage errors, nonresponse, or other selection effects, as noted in Professor Valliant's paper. In such cases, the MA estimator is doubly robust by construction: approximately unbiased if either the y-model $\mu(\cdot)$ or the π-model is correctly specified.

Specification of a linear working model $\mu(\boldsymbol{x}_k) = \boldsymbol{x}_k^\top \boldsymbol{\beta}$ in the MA recipe leads to the class of generalized regression (GREG) estimators, with special cases that cover many classic approaches (separate and combined ratio, separate and

combined regression, poststratification, etc.). GREG can be written in a weighted form,

$$\begin{aligned}\mathrm{GREG}(y_k) &= \sum_{k\in s}\left\{\frac{1}{\pi_k}+\left(T_{\boldsymbol{x}}-\mathrm{HT}(\boldsymbol{x}_k)\right)^{\top}\left(\sum_{k\in s}\frac{\boldsymbol{x}_k\boldsymbol{x}_k^{\top}}{\pi_k}\right)^{-1}\frac{\boldsymbol{x}_k}{\pi_k}\right\}y_k\\ &= \sum_{k\in s}\omega_{ks}y_k,\end{aligned}$$

where the GREG weights $\{\omega_{ks}\}_{k\in s}$ do not depend on y and can be applied generically to any response variable. These multipurpose weights have the property that they are calibrated to the population $\boldsymbol{x}$-totals, $\mathrm{GREG}(\boldsymbol{x}_k^{\top}) = \sum_{k\in U}\boldsymbol{x}_k^{\top}$, so that any y-variables with approximate linear relationships with $\boldsymbol{x}$ should be well-estimated.

GREG yields these nice properties of "generic" weights under mild data requirements: the complete microdata $(\pi_k, \boldsymbol{x}_k^{\top}, y_k)$ for $k \in s$ but only totals $T_{\boldsymbol{x}} = \sum_{k\in U}\boldsymbol{x}_k$ for the population. What about for other MA estimators? What data are needed and can multipurpose weights be obtained?

It is useful to divide MA approaches other than linear (GREG) into two sets: (A) those that are nearly linear, up to the values of a few unknown parameters, and (B) all others, featuring strong nonlinearity or many unknown parameters (including algorithmic approaches).

Among the nearly linear MA estimators of set (A), GREG-like weights can be obtained once values for a small number of unknown parameters are plugged in. These could be smoothing parameters in nonparametric regression approaches (Breidt and Opsomer 2000; Breidt et al. 2005; Goga 2005) or variance parameters in linear mixed models (LMMs). To illustrate, the LMM working model with a single unknown parameter is $y_k = \boldsymbol{x}_k^{\top}\boldsymbol{\beta} + \boldsymbol{z}_k^{\top}\boldsymbol{b} + \varepsilon_k$, where $\boldsymbol{b} \sim (0, \lambda^{-2}\boldsymbol{Q})$, and $\boldsymbol{Q}$ is positive definite and known. Let $\boldsymbol{c}_k^{\top} = [\boldsymbol{x}_k^{\top}, \boldsymbol{z}_k^{\top}]$ and $\Lambda = \mathrm{blockdiag}(\boldsymbol{0}, \lambda^2\boldsymbol{Q}^{-1})$. Then, the MA estimator based on the LMM is

$$\begin{aligned}\mathrm{LMM}(y_k) &= \sum_{k\in s}\left\{\frac{1}{\pi_k}+\left(T_{\boldsymbol{c}}-\mathrm{HT}(\boldsymbol{c}_k)\right)^{\top}\left(\sum_{k\in s}\frac{\boldsymbol{c}_k\boldsymbol{c}_k^{\top}}{\pi_k}+\Lambda\right)^{-1}\frac{\boldsymbol{c}_k}{\pi_k}\right\}y_k\\ &= \sum_{k\in s}\omega_{ks}y_k.\end{aligned}$$

Once λ is specified, the LMM weights $\{\omega_{ks}\}_{k\in s}$ are completely determined and can be applied to any y. Like the GREG, the LMM requires complete microdata $(\pi_k, \boldsymbol{x}_k^{\top}, \boldsymbol{z}_k^{\top}, y_k)$ for $k \in s$ but only totals $T_{\boldsymbol{c}} = \sum_{k\in U}[\boldsymbol{x}_k^{\top}, \boldsymbol{z}_k^{\top}]$ for the population.

Options for the parameters for a nearly linear approach include choices that are highly tuned to a specific y-variable of interest, or some compromise among a set of interesting ys, or a choice based on some criterion such as

penalization. In the single-parameter case above, for example, λ can be interpreted as a penalty. The covariates $\boldsymbol{x}_k$ are unpenalized, so the calibration $\text{LMM}(\boldsymbol{x}_k) = T_{\boldsymbol{x}}$ holds, but $\text{LMM}(z_k) \neq T_z$ due to the penalization. If $\lambda \to 0$, then there is no penalty, the LMM reverts to GREG on $\boldsymbol{c}_k$, and $\text{LMM}(z_k) \to T_z$. The LMM approach is sufficiently broad to cover ridge calibration (Beaumont and Bocci 2008), in which penalization relaxes calibration constraints, and penalized splines (Breidt et al. 2005), in which penalization equates to the degrees of freedom for smoothness of the approximating function.

Examples of MA approaches in set (B) include those based on generalized linear models and other parametric methods (Lehtonen and Veijanen 1998; Kennel and Valliant 2021), neural nets (Montanari and Ranalli 2005), single-index models (Wang 2009), generalized additive models (Opsomer et al. 2007), semiparametric additive models (Breidt et al. 2007), nonparametric additive models (Wang and Wang 2011), LASSO (McConville et al. 2017), and tree-based methods (Toth and Eltinge 2011; McConville and Toth 2019; Dagdoug et al. 2023). In some of these cases, special-purpose approximations can be employed to obtain multipurpose weights (e.g., McConville et al. 2017). Often, however, the model calibration method of Wu and Sitter (2001) is used to obtain weights, by using GREG with model predictions as the covariates. The model calibration method could use one model to predict one y, or multiple models to predict one y, or multiple models to predict multiple ys, if constructing a compromise set of weights.

As usual, complete microdata $(\pi_k, \boldsymbol{x}_k^\top, y_k)$ for $k \in s$ are required for set (B) methods. Unlike linear or nearly linear cases, complete auxiliary data $\{\boldsymbol{x}_k\}_{k \in U}$ for the population are also needed. Summary totals will not suffice. Though complete auxiliary data are required, it is not necessary to match the auxiliary data for the sample to the auxiliary data for the population; that is, the model-based prediction (summed over the population) and the design-bias adjustment (summed over the sample) in the MA form can be computed entirely separately, which is sometimes useful in practice.

In the linear case, GREG weights do not depend on any particular response y except in the choice of covariates. In the nearly linear cases (A), the weights depend on y through the choice of covariates and also possibly through the estimation or selection of the tuning parameters. The remaining cases (B) are more y-specific, as they may depend on the estimation of more parameters as well as on model-based predictions of y, if using model calibration. Nonetheless, multipurpose weights are available in each case.

The data needs and availability of multipurpose weights may not be entirely clear to practitioners, and this lack of clarity plus inertia might lead to underutilization of available MA techniques. Default methods are often some type of raking to population counts, and this default does not always reflect limitations of available control data.

While the emphasis in this discussion has been MA estimation, which uses flexible models and methods robust to model misspecification to take advantage of auxiliary information, similar ideas apply in other uses of models in surveys. Methods must be rigorously stress-tested offline prior to production mode. Academic researchers can do this to some extent using artificial data-generating mechanisms that are completely unlike the model assumed for the methodological development. But practitioners can always help out by creating test challenges, using real data directly or to build simulation engines for generating test data with realistic scale and complexity. A recent example is Benoit-Bryan and Mulrow (2021), who simulated replicate probability and nonprobability samples using data from a real study called Culture and Community in a Time of Crisis, which evaluated behaviors and attitudes during the global COVID-19 pandemic. The data are available via the Open Science Framework (https://osf.io/ygpzm/).

Professor Valliant's review of the evolution of the use of models in survey sampling reflects the longstanding approach of the entire field: theory and practice generate creative ideas that are met with both openness and cautiousness. With this approach, the field will continue to evolve to address new challenges.

REFERENCES

Beaumont, J. F., and Bocci, C. (2008), "Another Look at Ridge Calibration," *Metron*, 66, 5–20.

Benoit-Bryan, J., and Mulrow, E. (2021), "Exploring Nonprobability Methods with Simulations from a Common Data Source: Culture and Community in a Time of Crisis," in Proceedings of the Section on Survey Research Methods, Alexandria, VA: American Statistical Association, pp. 1633–1639.

Breidt, F. J., Claeskens, G., and Opsomer, J. D. (2005), "Model-Assisted Estimation for Complex Surveys Using Penalised Splines," *Biometrika*, 92, 4 831–846.

Breidt, F. J., and Opsomer, J. D. (2000), "Local Polynomial Regression Estimators in Survey Sampling," *Annals of Statistics*, 28, 1026–1053.

———. (2017), "Model-Assisted Survey Estimation with Modern Prediction Techniques," *Statistical Science*, 32, 2 190–205.

Breidt, F. J., Opsomer, J. D., Johnson, A. A., and Ranalli, M. G. (2007), "Semiparametric Model-Assisted Estimation for Natural Resource Surveys," *Survey Methodology*, 33, 35–44.

Dagdoug, M., Goga, C., and Haziza, D. (2023), "Model-Assisted Estimation through Random Forests in Finite Population Sampling," *Journal of the American Statistical Association*, 118, 1234–1251.

Goga, C. (2005), "Variance Reduction in Surveys with Auxiliary Information: A Nonparametric Approach Involving Regression Splines," *Canadian Journal of Statistics*, 33, 163–180.

Kennel, T. L., and Valliant, R. (2021), "Multivariate Logistic-Assisted Estimators of Totals from Clustered Survey Samples," *Journal of Survey Statistics and Methodology*, 9, 4 856–890.

Lehtonen, R., and Veijanen, A. (1998), "Logistic Generalized Regression Estimators," *Survey Methodology*, 24, 51–56.

McConville, K. S., Breidt, F. J., Lee, T. C., and Moisen, G. G. (2017), "Model-Assisted Survey Regression Estimation with the Lasso," *Journal of Survey Statistics and Methodology*, 5, 2 131–158.

McConville, K. S., and Toth, D. (2019), "Automated Selection of Post-Strata Using a Model-Assisted Regression Tree Estimator," *Scandinavian Journal of Statistics*, 46, 2 389–413.

Montanari, G. E., and Ranalli, M. G. (2005), "Nonparametric Model Calibration Estimation in Survey Sampling," *Journal of the American Statistical Association*, 100, 472 1429–1442.

Opsomer, J. D., Breidt, F. J., Moisen, G. G., and Kauermann, G. (2007), "Model-Assisted Estimation of Forest Resources with Generalized Additive Models (with Discussion)," *Journal of the American Statistical Association*, 102, 400–409.

Toth, D., and Eltinge, J. L. (2011), "Building Consistent Regression Trees from Complex Sample Data," *Journal of the American Statistical Association*, 106, 496 1626–1636.

Wang, L. (2009), "Single-Index Model-Assisted Estimation in Survey Sampling," *Journal of Nonparametric Statistics*, 21, 4 487–504.

Wang, L., and Wang, S. (2011), "Nonparametric Additive Model-Assisted Estimation for Survey Data," *Journal of Multivariate Analysis*, 102, 1126–1140.

Wu, C. F. J., and Sitter, R. R. (2001), "A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data," *Journal of the American Statistical Association*, 96, 185–193.