Discussion of "The Evolution of the Use of Models in Survey Sampling"

Jay Breidt
NORC at the University of Chicago

2022 Hansen Lecture: Richard Valliant

November 16, 2022

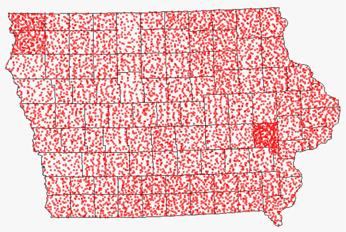
Thanks!

- Many thanks to the organizers and supporters of the Hansen lecture
  - ... for recruiting an outstanding Hansen lecturer
  - ... for inviting me to participate
  - ... for inspiring me to read Olkin's interview with Hansen in *Statistical Science* (1987)
- Thanks to Rick Valliant for an exceptional lecture
  - excellent exposition, expected given the papers and books we've all referenced
  - great reminder of the important role of models in surveys
  - trip down memory lane for me: my own evolution in understanding and use of models in surveys

My Three-Part Intro to Surveys

1. In graduate school, **took one sampling course** out of Cochran (1979, 3rd ed.) *Sampling Techniques*
   - mentions superpopulations, but not early and not often
   - course emphasis on derivations, not applications
2. **Taught several sampling courses** beginning at Iowa State University in 1991
   - first, out of Cochran (1979, 3rd edition)
   - subsequently, out of Särndal, Swensson, and Wretman (1992) *Model Assisted Survey Sampling*
3. **On-the-job training** in applied surveys
   - member of the Survey Section of the Iowa State Stat Lab
   - most of the work centered on USDA National Resources Inventory

USDA National Resources Inventory



- 300K PSUs in stratified two-stage sample
- longitudinal study of land cover and use, emphasis on soil erosion: loads of y-variables
- information at landscape, PSU and SSU levels
- "5% of the cases take 95% of the effort"
- need for generic, well-behaved weights

Model-Assisted Estimation a la SSW

- Model-assisted generalized regression (GREG) estimator introduces a **working model**

$$y_k = \mu(\boldsymbol{x}_k) + \epsilon_k = \boldsymbol{x}_k^T \boldsymbol{\beta} + \epsilon_k, \quad \epsilon_k \sim (0, \sigma^2)$$

- If the entire population were observed, use a standard statistical method to estimate $\mu(\cdot)$ by $m_N(\cdot)$:

$$m_N(\boldsymbol{x}_k) = \boldsymbol{x}_k^T \boldsymbol{B}_N = \boldsymbol{x}_k^T \left(\sum_{j \in U} \boldsymbol{x}_j \boldsymbol{x}_j^T \right)^{-1} \sum_{j \in U} \boldsymbol{x}_j y_j$$

- Since only a sample is observed, estimate $m_N(\cdot)$ by $\widehat{m_N}(\cdot)$:

$$\widehat{m_N}(\boldsymbol{x}_k) = \boldsymbol{x}_k^T \widehat{\boldsymbol{B}_N} = \boldsymbol{x}_k^T \left(\sum_{j \in s} \frac{\boldsymbol{x}_j \boldsymbol{x}_j^T}{\pi_j} \right)^{-1} \sum_{j \in s} \frac{\boldsymbol{x}_j y_j}{\pi_j}$$

Generalized Regression Estimation, continued

- Plug into model-assisted estimator form:

$$\begin{aligned}\text{GREG}(y_k) &= \sum_{k \in U} \mathbf{x}_k^T \widehat{\mathbf{B}_N} + \sum_{k \in s} \frac{y_k - \mathbf{x}_k^T \widehat{\mathbf{B}_N}}{\pi_k} \\ &= (\text{model-based prediction}) + (\text{design bias adjustment})\end{aligned}$$

  - classical survey **ratio estimator** and its variants
  - classical survey **regression estimator** and its variants
  - **post-stratification estimator**
  - ...
- Asymptotically design-unbiased and consistent even if the model is misspecified
- Smaller variance than HT if model is reasonably specified

GREG Produces Calibrated Weights

- GREG can also be written in weighted form:

$$
\begin{aligned}
\text{GREG}(y_k) &= \sum_{k \in U} \boldsymbol{x}_k^T \widehat{\boldsymbol{B}_N} + \sum_{k \in s} \frac{y_k - \boldsymbol{x}_k^T \widehat{\boldsymbol{B}_N}}{\pi_k} \\
&= \sum_{k \in s} \left\{ \frac{1}{\pi_k} + (t_{\boldsymbol{x}} - \text{HT}(\boldsymbol{x}_k))^T \left(\sum_{k \in s} \frac{\boldsymbol{x}_k \boldsymbol{x}_k^T}{\pi_k} \right)^{-1} \frac{\boldsymbol{x}_k}{\pi_k} \right\} y_k \\
&= \sum_{k \in s} \omega_{ks} y_k
\end{aligned}
$$

- GREG weights $\{\omega_{ks}\}$ do not depend on y and can be applied **generically** to any response variable
- GREG weights $\{\omega_{ks}\}$ are **calibrated** to the $\boldsymbol{X}$-totals:

$$
\text{GREG}(\boldsymbol{x}_k^T) = t_{\boldsymbol{x}}^T
$$

Information for Model-Assisted Estimation

- **Sample data:** covariates $\{\boldsymbol{x}_k\}$ and design weights $\{\pi_k^{-1}\}$ (no need to match to population)
- **Basic tabulations** available for the population
  - counts for categories
  - sums or means for continuous variables
  - suffices for additive models with untransformed covariates
- **Custom tabulations** available for the population, $\sum_{k \in U} \boldsymbol{h}(\boldsymbol{x}_k)$, for known transformations, $\boldsymbol{h}(\cdot)$
  - polynomials or other transformations of continuous variables, including spline basis functions
  - interactions, including continuous by categorical
- **Complete microdata** $\{\boldsymbol{x}_k\}_{k \in U}$ for all population elements

General Recipe for Model-Assisted Estimation

- Specify a working model, $y_k = \mu(\boldsymbol{x}_k) + \epsilon_k$, $\epsilon_k \sim (0, \sigma^2)$
- Write down infeasible full-population "estimator," $m_N(\cdot)$
- Create feasible survey-weighted version, $\widehat{m_N}(\cdot)$
- Plug in and write **model-assisted estimator** as

$$\begin{aligned}\mathrm{MA}(y_k) &= \sum_{k \in U} \widehat{m_N}(\boldsymbol{x}_k) + \sum_{k \in s} \frac{y_k - \widehat{m_N}(\boldsymbol{x}_k)}{\pi_k} \\ &= (\text{model-based prediction}) + (\text{design bias adjustment})\end{aligned}$$

  - good properties like those of GREG under mild conditions
  - **doubly-robust** by construction if π_k must be estimated
- But what about "generic" weights?
  - depends on whether $\widehat{m_N}(\cdot)$ is **really linear** (GREG), **sort-of linear**, or **not really linear**

MA Estimation with “Sort-of Linear” Methods

- **Sort-of linear:** linear except for a few unknown parameters
  - GREG-like weights once parameter values are plugged in
- Unknown smoothing parameters in **nonparametric regression**
  - local polynomial regression (Breidt and Opsomer 2000)
  - regression splines (Goga 2005)
  - penalized splines (Breidt, Claeskens, Opsomer 2005)
- Unknown variance parameters in **linear mixed models**
  - ridge calibration (Beaumont and Bocci 2008)
  - penalized splines
- **Options** for choosing parameters?
  - highly tuned to specific y
  - compromise among interesting y's
  - penalization or other criteria

MA Estimation with Linear Mixed Model

- LMM working model: $y_k = \boldsymbol{x}_k^T\beta + \boldsymbol{z}_k^T\boldsymbol{b} + \epsilon_k$, $\boldsymbol{b} \sim (\boldsymbol{0}, \lambda^{-2}\boldsymbol{Q})$
- Let $\boldsymbol{c}_k^T = [\boldsymbol{x}_k^T, \boldsymbol{z}_k^T]$ and $\boldsymbol{\Lambda} = \text{blockdiag}(\boldsymbol{0}, \lambda^2\boldsymbol{Q}^{-1})$

$$\begin{aligned}\text{LMM}(y_k) &= \sum_{k\in U} \boldsymbol{c}_k^T\widehat{\boldsymbol{B}_N} + \sum_{k\in s}\frac{y_k - \boldsymbol{c}_k^T\widehat{\boldsymbol{B}_N}}{\pi_k} \\ &= \sum_{k\in s}\left\{\frac{1}{\pi_k} + (t_{\boldsymbol{c}} - \text{HT}(\boldsymbol{c}_k))^T\left(\sum_{k\in s}\frac{\boldsymbol{c}_k\boldsymbol{c}_k^T}{\pi_k} + \boldsymbol{\Lambda}\right)^{-1}\frac{\boldsymbol{c}_k}{\pi_k}\right\}y_k\end{aligned}$$

- $\text{LMM}(\boldsymbol{x}_k^T) = t_{\boldsymbol{x}}^T$, but $\text{LMM}(\boldsymbol{z}_k^T) \neq t_{\boldsymbol{z}}^T$, due to the penalization
    - $\lambda \to \infty$ implies GREG on $\boldsymbol{x}_k$ only
    - $\lambda \to 0$ implies GREG on $(\boldsymbol{x}_k, \boldsymbol{z}_k)$

MA Estimation with "Not Really Linear" Methods

- **Not really linear:** many unknown parameters, algorithmic approaches
  - generalized linear models, other parametric methods (Lehtonen and Veijanen 1998, Kennel and Valliant 2021)
  - neural nets (Montanari and Ranalli 2005), single-index models (Wang 2009)
  - additive models: generalized (Opsomer et al. 2007), semiparametric (Breidt et al. 2007), nonparametric (Wang and Wang 2011)
  - selection and shrinkage methods (McConville et al. 2017)
  - tree-based methods (Toth and Eltinge 2011, McConville and Toth 2019, Dagdoug et al. 2021, 2022)
- Most use **model calibration** of Wu and Sitter (2001) to obtain weights
  - GREG with model predictions as covariates

Dependence of Weights on y

- **Really linear:** GREG weights

$$\left\{ \frac{1}{\pi_k} + (t_{\boldsymbol{x}} - \text{HT}(\boldsymbol{x}_k))^T \left(\sum_{k \in s} \frac{\boldsymbol{x}_k \boldsymbol{x}_k^T}{\pi_k} \right)^{-1} \frac{\boldsymbol{x}_k}{\pi_k} \right\}$$

  do not depend on y except through choice of covariates

- **Sort-of linear:** MA weights depend on
  - choice of covariates, as with GREG
  - estimation/selection of tuning parameters, usually a small number
- **Not really linear:** MA weights depend on
  - choice of covariates, as with GREG
  - estimation/selection of parameters, possibly a large number
  - model-based predictions of y, if using model calibration

Final Thoughts on Models in Surveys

- Emphasis here on models used to take advantage of auxiliary information in **model-assisted estimation**
  - flexible models and methods robust to model misspecification
  - similar ideas apply in other uses of models in surveys
- Models are **extremely useful**
  - for organizing and communicating thoughts
  - for deriving estimators with good properties
  - for assessing expected behavior under ideal conditions
  - for identifying non-ideal conditions
- We should **maintain healthy skepticism** of models while being open to new ideas
  - robustness is essential in production environments
  - researchers should test methods with data generating mechanisms completely unlike the assumed model
  - practitioners could create test challenges, real or deep-fake