Discussion of “The Evolution of the Use of Models in Survey Sampling”

Jay Breidt

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 generalized regression (GREG) estimator introduces a **working model**

\[ y_k = \mu(x_k) + \epsilon_k = x_k^T \beta + \epsilon_k, \quad \epsilon_k \sim (0, \sigma^2) \]

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

\[
m_N(x_k) = x_k^T B_N = x_k^T \left( \sum_{j \in U} x_j x_j^T \right)^{-1} \sum_{j \in U} x_j y_j
\]

• Since only a sample is observed, estimate \( m_N(\cdot) \) by \( \hat{m}_N(\cdot) \):

\[
\hat{m}_N(x_k) = x_k^T \hat{B}_N = x_k^T \left( \sum_{j \in s} \frac{x_j x_j^T}{\pi_j} \right)^{-1} \sum_{j \in s} \frac{x_j y_j}{\pi_j}
\]
Generalized Regression Estimation, continued

• Plug into model-assisted estimator form:

\[
GREG(y_k) = \sum_{k \in U} x_k^T \hat{B}_N + \sum_{k \in s} \frac{y_k - x_k^T \hat{B}_N}{\pi_k}
\]

= (model-based prediction) + (design bias adjustment)

• 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:

\[
GREG(y_k) = \sum_{k \in U} x_k^T \hat{B}_N + \sum_{k \in s} \frac{y_k - x_k^T \hat{B}_N}{\pi_k}
\]

\[
= \sum_{k \in s} \left\{ \frac{1}{\pi_k} + (t_x - HT(x_k))^T \left( \sum_{k \in s} \frac{x_k x_k^T}{\pi_k} \right)^{-1} \frac{x_k}{\pi_k} \right\} y_k
\]

\[
= \sum_{k \in s} \omega_{ks} y_k
\]

- GREG weights \( \{\omega_{ks}\} \) do not depend on \( y \) and can be applied \textbf{generically} to any response variable.

- GREG weights \( \{\omega_{ks}\} \) are \textbf{calibrated} to the \( X \)-totals:

\[
GREG(x_k^T) = t_x^T
\]
• **Sample data:** covariates \( \{x_k\} \) 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, \( \sum_{k \in U} h(x_k) \), for known transformations, \( h(\cdot) \)
  - polynomials or other transformations of continuous variables, including spline basis functions
  - interactions, including continuous by categorical

• **Complete microdata** \( \{x_k\}_{k \in U} \) for all population elements
General Recipe for Model-Assisted Estimation

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

\[
\text{MA}(y_k) = \sum_{k \in U} \hat{m}_N(x_k) + \sum_{k \in s} \frac{y_k - \hat{m}_N(x_k)}{\pi_k}
= \text{(model-based prediction)} + \text{(design bias adjustment)}
\]

- good properties like those of GREG under mild conditions
- **doubly-robust** by construction if \( \pi_k \) must be estimated
- But what about "generic" weights?
  - depends on whether \( \hat{m}_N(\cdot) \) 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: $y_k = x_k^T \beta + z_k^T b + \epsilon_k$, $b \sim (0, \lambda^{-2} Q)$
- Let $c_k^T = [x_k^T, z_k^T]$ and $\Lambda = \text{blockdiag}(0, \lambda^2 Q^{-1})$

$$
\text{LMM}(y_k) = \sum_{k \in U} c_k^T \hat{B}_N + \sum_{k \in s} \frac{y_k - c_k^T \hat{B}_N}{\pi_k} \\
= \sum_{k \in s} \left\{ \frac{1}{\pi_k} + (t_c - \text{HT}(c_k))^T \left( \sum_{k \in s} \frac{c_k c_k^T}{\pi_k} + \Lambda \right)^{-1} \frac{c_k}{\pi_k} \right\} y_k
$$

- LMM($x_k^T$) = $t_x^T$, but LMM($z_k^T$) $\neq t_z^T$, due to the penalization
  - $\lambda \to \infty$ implies GREG on $x_k$ only
  - $\lambda \to 0$ implies GREG on $(x_k, z_k)$
• **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
• **Really linear:** GREG weights

\[
\left\{ \frac{1}{\pi_k} + \left( t_x - HT(x_k) \right)^T \left( \sum_{k \in s} \frac{x_k x_k^T}{\pi_k} \right)^{-1} \frac{x_k}{\pi_k} \right\}
\]

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