Model-assisted survey regression estimation with the lasso [1]

F. Jay Breidt
Colorado State University

Opening Workshop on Computational Methods in Social Sciences
SAMSI August 2013

This research was supported in part by the US National Science Foundation (SES0922142).

Model-assisted survey regression estimation with the lasso

- (Model-assisted survey regression estimation):
  - joint work with Jean Opsomer, Colorado State, and various other colleagues
  - bringing in nonparametric and semiparametric regression methods into classical survey statistics
- (with the lasso):
  - joint work with former PhD student Kelly McConville, Whitman College; Thomas Lee, UC-Davis; and Gretchen Moisen, US Forest Service

Finite population and probability sampling

- Finite population $U = \{1, 2, \ldots, N\}$
- Response variable y_k, $k \in U$
  - these are just unknown real numbers, with no probability structure
- Goal: estimate finite population total

$$t_y = \sum_{k \in U} y_k$$

- Draw probability sample $s \subset U$ with $\Pr[k \in s] = \pi_k > 0$

- Sample membership indicator $I_k = 1$ if $k \in s$, $I_k = 0$ otherwise

$$\mathsf{E}\left[I_k\right] = \pi_k, \text{ averaging over all possible samples}$$

- Use this repeated-sampling probability structure for statistical inference
- Unbiased Horvitz-Thompson estimator of t_y is

$$\mathsf{HT}(y_k) = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k I_k}{\pi_k}$$

Variance of the Horvitz-Thompson estimator

- Depends on covariance structure of $\{I_k\}_{k \in U}$:

$$\begin{aligned} \text{Var}\left(\sum_{k \in U} y_k \frac{I_k}{\pi_k}\right) &= \sum_{k,\ell \in U} \text{Cov}\left(I_k, I_\ell\right) \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} \\ &= \sum_{k,\ell \in U} \Delta_{k\ell} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} \end{aligned}$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$ and $\pi_{k\ell} = \text{Pr}\left[I_k = 1, I_\ell = 1\right]$

- Provided $\pi_{k\ell} > 0$ for all $k, \ell \in U$,

$$\widehat{V}(\mathsf{HT}) = \sum_{k,\ell \in U} \Delta_{k\ell} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} \frac{I_k I_\ell}{\pi_{k\ell}}$$

  is unbiased for

$$\mathsf{Var}\,(\mathsf{HT}) = \sum_{k,\ell \in U} \Delta_{k\ell} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell}$$

- Under very mild design conditions,
  - $\widehat{V}(\mathrm{HT})$ is consistent for $\mathrm{Var}\,(\mathrm{HT})$
  - HT is asymptotically normal and confidence intervals can be based on this fact in moderate to large samples

$$\left\{\widehat{V}(\mathrm{HT})\right\}^{-1/2}\left\{\mathrm{HT}(y_k) - t_y\right\} \xrightarrow{\mathcal{L}} \mathcal{N}\,(0,1)$$

- Involves only "finite second moments" of $\{y_k\}$,
  - not distributional assumptions
  - not dependence assumptions
- Good design makes normal approximations better

- Suppose we have auxiliary information vector $\boldsymbol{x}_k$, $k \in U$
- Also have a "method" $m(\cdot)$ for predicting y_k from $\boldsymbol{x}_k$:

$$y_k \simeq m(\boldsymbol{x}_k)$$

  - method $m(\cdot)$ does not depend on the sample
  - e.g., inflation-adjust an old census value

- Unbiased difference estimator of t_y is then

$$\text{Diff}_m(y_k) = \sum_{k \in U} m(\boldsymbol{x}_k) + \sum_{k \in s} \frac{y_k - m(\boldsymbol{x}_k)}{\pi_k}$$

Variance of the difference estimator

$$\text{Var}\left(\sum_{k\in U} m(\boldsymbol{x}_k) + \sum_{k\in U}(y_k - m(\boldsymbol{x}_k))\frac{I_k}{\pi_k}\right)$$
$$= \sum_{k,\ell\in U} \Delta_{k\ell}\frac{y_k - m(\boldsymbol{x}_k)}{\pi_k}\frac{y_\ell - m(\boldsymbol{x}_\ell)}{\pi_\ell}$$

- Compare to Horvitz-Thompson estimator:

$$\text{Var}\left(\sum_{k\in U} y_k\frac{I_k}{\pi_k}\right) = \sum_{k,\ell\in U}\Delta_{k\ell}\frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell}$$

Variance estimator for the difference estimator

- Provided $\pi_{k\ell} > 0$ for all $k, \ell \in U$,

$$\widehat{V}\left(\text{Diff}_m\right) = \sum_{k,\ell \in U} \Delta_{k\ell} \frac{y_k - m(\boldsymbol{x}_k)}{\pi_k} \frac{y_\ell - m(\boldsymbol{x}_\ell)}{\pi_\ell} \frac{I_k I_\ell}{\pi_{k\ell}}$$

  is unbiased for

$$\text{Var}\left(\text{Diff}_m\right) = \sum_{k,\ell \in U} \Delta_{k\ell} \frac{y_k - m(\boldsymbol{x}_k)}{\pi_k} \frac{y_\ell - m(\boldsymbol{x}_\ell)}{\pi_\ell}$$

- Inherits asymptotic normality from HT under mild conditions.

Summary so far . . .

- Difference estimator is **exactly** unbiased, regardless of the quality of the method $m(\cdot)$

$$\text{Diff}_m(y_k) = \sum_{k \in U} m(\boldsymbol{x}_k) + \text{HT}\left(y_k - m(\boldsymbol{x}_k)\right)$$

- Has smaller variance than $\text{HT}(y_k)$ provided "residuals"

$$y_k - m(\boldsymbol{x}_k)$$

  have smaller variation than "raw values" y_k
- Have an **exactly** unbiased variance estimator
- Results require that $m(\cdot)$ does not depend on the sample

- Difference estimator requires method $m(\cdot)$ independent of the sample
- Model-assisted estimator introduces a **working model**

$$y_k = \mu(\boldsymbol{x}_k) + \epsilon_k$$

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

Model-assisted estimation, continued

- Since only a sample is observed, estimate $m_N(\cdot)$ by $\widehat{m}_N(\cdot)$
  - not independent of the sample
- Plug $\widehat{m}_N(\cdot)$ into the difference estimator form:

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

- Plug $\widehat{m}_N(\cdot)$ into the variance estimator:

$$\widehat{V}\left(\mathsf{MA}\right) = \sum_{k,\ell \in U} \Delta_{k\ell} \frac{y_k - \widehat{m}_N(\boldsymbol{x}_k)}{\pi_k} \frac{y_\ell - \widehat{m}_N(\boldsymbol{x}_\ell)}{\pi_\ell} \frac{I_k I_\ell}{\pi_{k\ell}}$$

Important case: Generalized regression estimation

- Working model is heteroskedastic multiple regression:

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

- If the entire population were observed, use weighted least squares:

$$m_N(\boldsymbol{x}_k) = \boldsymbol{x}_k'\boldsymbol{B}_N = \boldsymbol{x}_k' \left(\sum_{j \in U} \frac{\boldsymbol{x}_j \boldsymbol{x}_j'}{\sigma_j^2} \right)^{-1} \sum_{j \in U} \frac{\boldsymbol{x}_j y_j}{\sigma_j^2}$$

Generalized regression estimation, continued

- Estimate finite population fit from the observed sample:

$$\widehat{m}_N(\boldsymbol{x}_k) = \boldsymbol{x}_k' \widehat{\boldsymbol{B}_N} = \boldsymbol{x}_k' \left(\sum_{j \in s} \frac{\boldsymbol{x}_j \boldsymbol{x}_j'}{\pi_j \sigma_j^2} \right)^{-1} \sum_{j \in s} \frac{\boldsymbol{x}_j y_j}{\pi_j \sigma_j^2}$$

- $\widehat{\boldsymbol{B}_N}$ is asymptotically design unbiased and consistent for $\boldsymbol{B}_N$ regardless of the quality of the working model specification

Generalized regression estimation, continued

- Plug into model-assisted form:

$$\text{GREG}(y_k) = \sum_{k \in U} \boldsymbol{x}_k' \widehat{\boldsymbol{B}_N} + \sum_{k \in s} \frac{y_k - \boldsymbol{x}_k' \widehat{\boldsymbol{B}_N}}{\pi_k}$$

- Plug into the variance estimator:

$$\widehat{V}\left(\text{GREG}\right) = \sum_{k,\ell \in U} \Delta_{k\ell} \frac{y_k - \boldsymbol{x}_k' \widehat{\boldsymbol{B}_N}}{\pi_k} \frac{y_\ell - \boldsymbol{x}_\ell' \widehat{\boldsymbol{B}_N}}{\pi_\ell} \frac{I_k I_\ell}{\pi_{k\ell}}$$

GREG examples

- Classical survey ratio estimator: x_k is scalar, model is heteroskedastic regression through the origin
- Classical survey regression estimator: x_k is scalar, model is homoskedastic simple linear regression
- Post-stratification estimator: $\boldsymbol{x}_k$ is vector of indicators for categorical covariate
- ...

Properties of GREG

- Write GREG as

$$
\begin{aligned}
\text{GREG}(y_k) &= \sum_{k\in U} \boldsymbol{x}_k' \widehat{\boldsymbol{B}_N} + \sum_{k\in s} \frac{y_k - \boldsymbol{x}_k' \widehat{\boldsymbol{B}_N}}{\pi_k} \\
&= \sum_{k\in U} \boldsymbol{x}_k' \boldsymbol{B}_N + \sum_{k\in U} \frac{(y_k - \boldsymbol{x}_k' \boldsymbol{B}_N) I_k}{\pi_k} \\
&\quad + \left(\widehat{\boldsymbol{B}_N} - \boldsymbol{B}_N\right)' \sum_{k\in U} \boldsymbol{x}_k \left(1 - \frac{I_k}{\pi_k}\right) \\
&= \text{Diff}_{m_N}(y_k) + (\text{smaller-order term})
\end{aligned}
$$

GREG estimator behaves like difference estimator

- Asymptotically unbiased (and mean square consistent), regardless of the quality of the working model $\mu(\cdot)$
- Variance is asymptotically equivalent to

$$\begin{aligned}
&\text{Var}\left(\sum_{k\in U}\boldsymbol{x}_k'\boldsymbol{B}_N+\sum_{k\in U}(y_k-\boldsymbol{x}_k'\boldsymbol{B}_N)\frac{I_k}{\pi_k}\right)\\
&=\sum_{k,\ell\in U}\Delta_{k\ell}\frac{y_k-\boldsymbol{x}_k'\boldsymbol{B}_N}{\pi_k}\frac{y_\ell-\boldsymbol{x}_k'\boldsymbol{B}_N}{\pi_\ell}
\end{aligned}$$

- Smaller asymptotic variance than $\text{HT}(y_k)$ provided residuals $y_k-\boldsymbol{x}_k'\boldsymbol{B}_N$ have less variation than raw values y_k

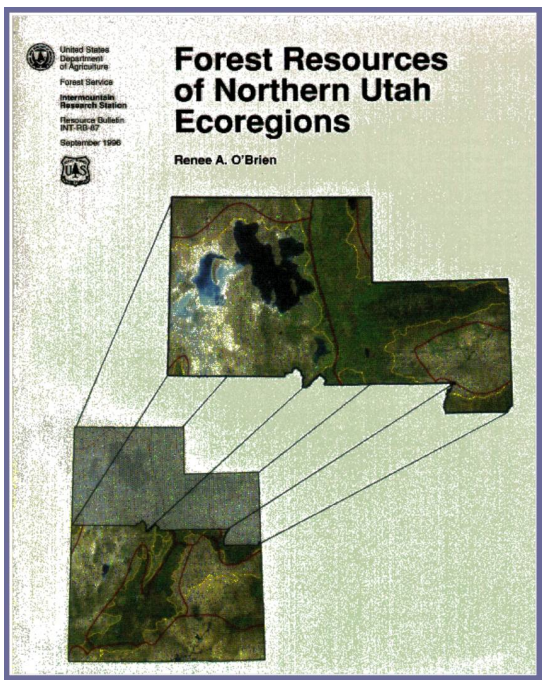
- GREG can also be written in weighted form:

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

- The GREG weights $\{\omega_{ks}\}$ do not depend on y and can be applied to any response variable

US Forest Inventory and Analysis: Many y's!

- Estimates required for forest area, wood volume, growth, mortality,...
- By region, species and other classifications



- Note that the weights $\{\omega_{ks}\}$ are calibrated to the $\boldsymbol{X}$-totals:

$$\begin{aligned}
\mathsf{GREG}(\boldsymbol{x}'_k) &= \sum_{k\in s}\left\{\frac{1}{\pi_k} + (t_{\boldsymbol{x}} - \mathsf{HT}(\boldsymbol{x}_k))'\left(\sum_{k\in s}\frac{\boldsymbol{x}_k\boldsymbol{x}'_k}{\pi_k}\right)^{-1}\frac{\boldsymbol{x}_k}{\pi_k}\right\}\boldsymbol{x}'_k \\
&= \mathsf{HT}(\boldsymbol{x}'_k) + (t_{\boldsymbol{x}} - \mathsf{HT}(\boldsymbol{x}_k))'\left(\sum_{k\in s}\frac{\boldsymbol{x}_k\boldsymbol{x}'_k}{\pi_k}\right)^{-1}\sum_{k\in s}\frac{\boldsymbol{x}_k\boldsymbol{x}'_k}{\pi_k} \\
&= t'_{\boldsymbol{x}}
\end{aligned}$$

- GREG will be very efficient if y_k is approximately a linear combination of $\boldsymbol{x}_k$

GREG calibration features

- GREG$(\boldsymbol{x}'_k) = t'_{\boldsymbol{x}}$
- Calibration reproduces known population information
  - internal consistency across statistical system
  - reassuring for users, logistically convenient for agencies
- GREG weight adjustments may be large
  - extreme weights, negative weights are possible
  - many methods developed to trim or stabilize weights, including ridge calibration (Rao and Singh 1997, *ASA Proc.*; Park and Fuller 2005, *Survey Methodology*; Montanari and Ranalli 2009, Stats Canada workshop.)

Alternatives to GREG

- Keep the calibration, change the metric: Deville and Särndal (1992, *JASA*) minimize $d\left(\pi_k^{-1}, \omega_{ks}\right)$ subject to $\sum_{k \in s} \omega_{ks} \boldsymbol{x}_k = t_{\boldsymbol{x}}$
- Specify the working model more flexibly:
  - Local polynomial regression (Breidt and Opsomer 2000)
  - Neural nets (Montanari and Ranalli 2005)
  - Penalized splines (Breidt, Claeskens, Opsomer 2005); Regression splines (Goga 2005)
  - Generalized additive models (Opsomer, Breidt, Moisen, and Kauermann 2007); Nonparametric additive models (Wang and Wang 2011)

General recipe for model-assisted estimation

- Write down your favorite $m_N(\cdot)$ you would use if the entire population were observed
- Create survey-weighted version, $\widehat{m}_N(\cdot)$
- Plug in and write model-assisted estimator as

$$\begin{aligned}\text{MA}(y_k) &= \sum_{k\in U} \widehat{m}_N(\boldsymbol{x}_k) + \sum_{k\in s} \frac{y_k - \widehat{m}_N(\boldsymbol{x}_k)}{\pi_k} \\ &= \text{Diff}_{m_N}(y_k) + (\text{smaller-order term})\end{aligned}$$

Example: Local polynomial regression

- Breidt and Opsomer (2000), *Ann. Stat.*
- Working model: $\mu(\cdot)$ is a smooth function of scalar x
- Estimate $\mu(\cdot)$ via local polynomial regression:

$$m_N(x_i) = (1, 0, \ldots, 0) \left(\boldsymbol{X}'_{Ui} \boldsymbol{W}_{Ui} \boldsymbol{X}_{Ui} \right)^{-1} \boldsymbol{X}'_{Ui} \boldsymbol{W}_{Ui} \boldsymbol{y}_U$$

where

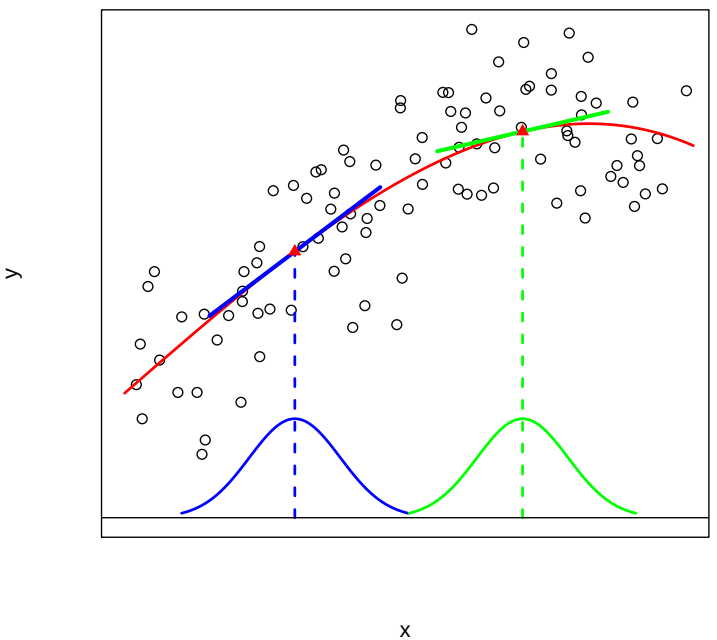
$$\boldsymbol{X}_{Ui} = \left[1 \;\; x_j - x_i \;\; \cdots \;\; (x_j - x_i)^q \right]_{j \in U}$$

and

$$\boldsymbol{W}_{Ui} = \text{diag} \left\{ \frac{1}{h} K \left(\frac{x_j - x_i}{h} \right) \right\}_{j \in U}$$

Example: Local linear regression

- Fit line locally (as defined by kernel weights) and read off the local intercept



Properties of the LPR estimator?

- Estimate $m_N(\cdot)$ using survey weights to get $\widehat{m}_N(\cdot)$, plug in to model-assisted form

$$\begin{aligned}
\text{LPR}(y_k) &= \sum_{k\in U} \widehat{m}_N(\boldsymbol{x}_k) + \sum_{k\in s} \frac{y_k - \widehat{m}_N(\boldsymbol{x}_k)}{\pi_k} \\
&= \sum_{k\in U} m_N(\boldsymbol{x}_k) + \sum_{k\in U} \frac{(y_k - m_N(\boldsymbol{x}_k))I_k}{\pi_k} \\
&\quad + \sum_{k\in U} (\widehat{m}_N(\boldsymbol{x}_k) - m_N(\boldsymbol{x}_k)) \left(1 - \frac{I_k}{\pi_k}\right) \\
&= \text{Diff}_{m_N}(y_k) + \text{(smaller-order term??)}
\end{aligned}$$

Asymptotic framework

- Consider sequence of finite populations with $n_N \to \infty$ as $N \to \infty$
- Smoothing assumptions:
  - kernel $K(\cdot)$ is symmetric, continuous, and compactly supported
  - bandwidth $h_N \to 0$ and $Nh_N^2 \to \infty$
- Design assumptions:
  - $\min_{i \in U_N} \pi_i \geq \lambda > 0$ and $\min_{i,j \in U_N} \pi_{ij} \geq \lambda^* > 0$
  - limited dependence:

$$\max_{i,j \in U_N : i \neq j} \quad |\pi_{ij} - \pi_i \pi_j| = O\left(n_N^{-1}\right)$$

$$\max_{(i,j,k,\ell) \text{ distinct}} \quad |\mathsf{E}\left[(I_i - \pi_i)(I_j - \pi_j)(I_k - \pi_k)(I_\ell - \pi_\ell)\right]| = O\left(N^{-2}\right)$$

$$\max_{(i,j,k,\ell) \text{ distinct}} \quad |\mathsf{E}\left[(I_i I_j - \pi_{ij})(I_k I_\ell - \pi_{k\ell})\right]| = o(1)$$

$$\max_{(i,j,k) \text{ distinct}} \quad |\mathsf{E}\left[(I_i - \pi_i)^2 (I_j - \pi_j)(I_k - \pi_k)\right]| = O\left(n_N^{-1}\right)$$

LPR estimator behaves like difference estimator

Under the above asymptotic framework,

- LPR estimator is mean square consistent for t_y
- Variance is asymptotically equivalent to

$$\sum_{k,\ell \in U} \Delta_{k\ell} \frac{y_k - m_N(\boldsymbol{x}_k)}{\pi_k} \frac{y_\ell - m_N(\boldsymbol{x}_\ell)}{\pi_\ell}$$

- Standard plug-in variance estimator is consistent
- Smaller asymptotic variance than HT provided residuals $y_k - m_N(\boldsymbol{x}_k)$ have less variation than raw values y_k

Weighting and calibration for LPR

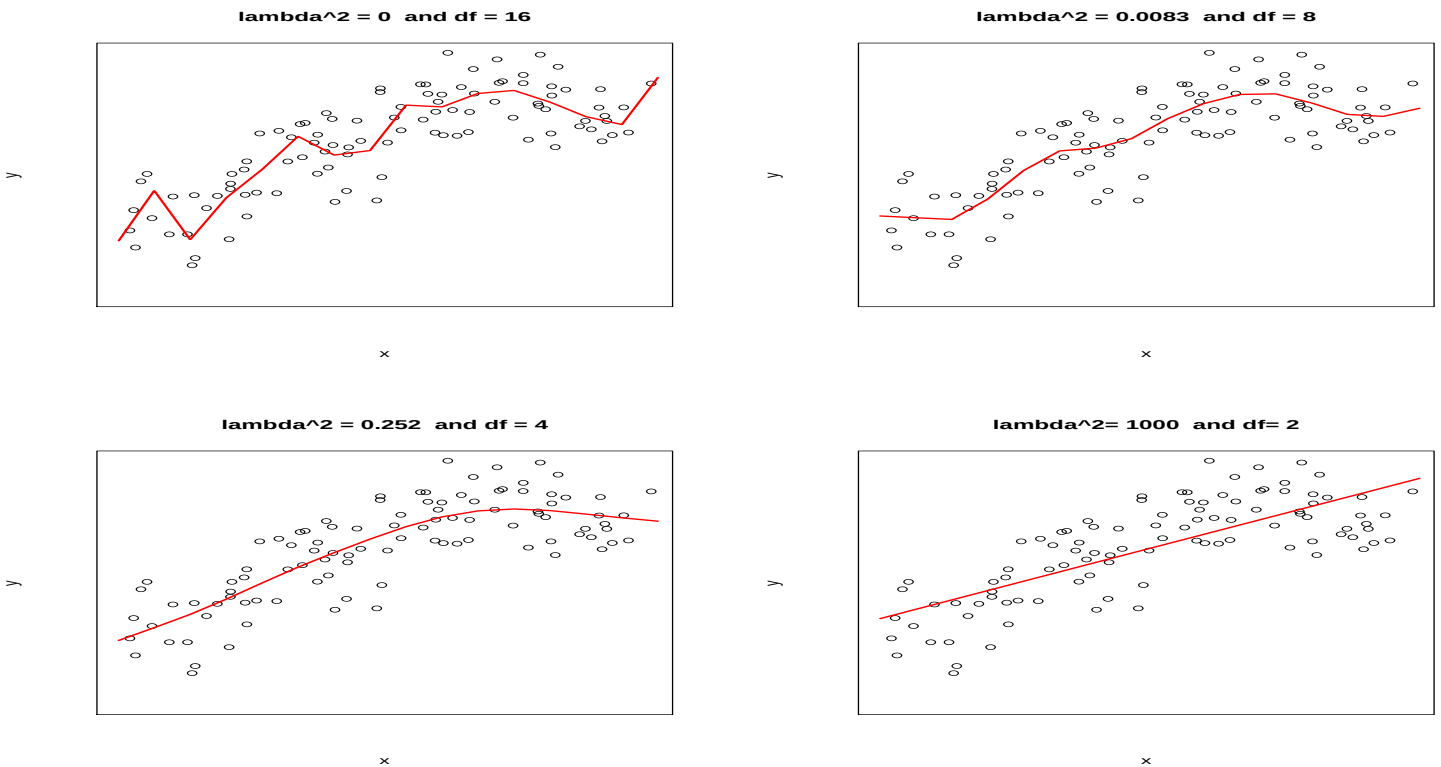
- LPR$(y_i) = \sum_{i \in s} \omega_{is} y_i$ with weights independent of y
- For qth order local polynomial, weights are calibrated to powers of x:

$$\sum_{i \in s} \omega_{is} x_i^\ell = \sum_{i \in U_N} x_i^\ell \quad (\ell = 0, 1, \ldots, q)$$

- LPR will be particularly effective if y is approximately a qth order polynomial in x

Example: Penalized spline regression

- K fixed, known knots and K basis functions
- Penalty parameter λ^2 determines degrees of freedom of the smooth



- Breidt, Claeskens and Opsomer (2005, $Biometrika$)
- Choose penalty λ^2 to give specified degrees of freedom
- Formulate p-spline as linear mixed model (LMM)
- Write down LMM fit for the entire finite population:

$$[m_N(\boldsymbol{c}_k)]_{k\in U} = \boldsymbol{C}(\boldsymbol{C}'\boldsymbol{C} + \boldsymbol{\Lambda})^{-1}\boldsymbol{C}'\boldsymbol{y}_U$$

  with $\boldsymbol{\Lambda} = \text{blockdiag}(\boldsymbol{0}, \lambda^2\boldsymbol{I}_K)$
- Estimate $m_N(\cdot)$ using survey weights to get $\widehat{m}_N(\cdot)$, plug in to model-assisted form

P-spline estimator behaves like difference estimator

- Under standard asymptotic framework with K fixed,
  - p-spline estimator is mean square consistent for t_y
  - variance is asymptotically equivalent to

$$\sum_{k,\ell \in U} \Delta_{k\ell} \frac{y_k - m_N(\boldsymbol{c}_k)}{\pi_k} \frac{y_\ell - m_N(\boldsymbol{c}_\ell)}{\pi_\ell}$$

  - standard plug-in variance estimator is consistent
  - smaller asymptotic variance than HT provided residuals $y_k - m_N(\boldsymbol{c}_k)$ have less variation than raw values y_k
- McConville and Breidt (2013, *J. Nonpar. Stat.*) prove above results with $K \to \infty$

P-spline weighting and calibration

- LMM model-assisted estimator (including p-spline) can be written

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

- Calibrated on $\boldsymbol{X}$, $\text{LMM}(\boldsymbol{x}_k') = t_{\boldsymbol{x}}'$, but not on $\boldsymbol{Z}$, $\text{LMM}(\boldsymbol{z}_k') \neq t_{\boldsymbol{z}}'$, due to the penalization

Comparing local polynomials with p-splines

- Both LPR and p-splines have good robustness properties:
  - comparable efficiency to GREG when parametric working model is correct
  - better efficiency when parametric working model is incorrect
  - better-behaved weights than GREG (e.g., almost never negative)
- P-splines extend much more readily than kernels to semi-parametric models:
  - additional $\boldsymbol{X}$ variables, continuous or categorical, with calibration
  - additional $\boldsymbol{Z}$ variables, continuous or categorical, without calibration

"Model Calibration": GREG with Model Predictions

- Both LPR and p-splines had linear structure, allowing calibrated y-independent weights
- Nonlinear methods are typically uncalibrated, but...
- Wu and Sitter (2001) *JASA*: regress y_k on $\widehat{m_N}(\boldsymbol{x}_k)$:

$$\begin{aligned}\mathsf{WS}(y_k) &= \sum_{k\in s}\frac{y_k-\widehat{m_N}(\boldsymbol{x}_k)\widehat{R}}{\pi_k}+\sum_{k\in U}\widehat{m_N}(\boldsymbol{x}_k)\widehat{R}\\ &= \sum_{k\in s}\left\{\frac{1}{\pi_k}+\left(\sum_{k\in U}\widehat{m_N}(\boldsymbol{x}_k)-\sum_{k\in s}\frac{\widehat{m_N}(\boldsymbol{x}_k)}{\pi_k}\right)\frac{\widehat{m_N}(\boldsymbol{x}_k)/\pi_k}{\sum_{k\in s}\widehat{m_N}(\boldsymbol{x}_k)^2/\pi_k}\right\}y_k\end{aligned}$$

yielding weights, but these depend on y

Forest Inventory and Analysis Program

- Conducted by United States Forest Service
- Nationwide network of 400,000 sample plots, visited every five years in rotating panels
- Goal: national and regional estimates of t_y or $\bar{y}_U$ for various y: forest area, wood volume, growth, mortality,...
  - expensive data y are field-collected or manually interpreted from aerial photography
  - interest in using cheap auxiliary data $\boldsymbol{x}$ from remote sensing and other spatial data sources

Auxiliary information for all $k \in U$ in real applications?

- Required auxiliary info comes from
  - "wall-to-wall" remote sensing, like satellite imagery
  - digital elevation models
  - geographic information system data products
- Working model is $\mu(\boldsymbol{x}_k) = \boldsymbol{x}_k'\boldsymbol{\beta}$, as with GREG:
  - but auxiliary variables may be highly correlated
  - may have poor predictive ability for response variables
- Consider using model selection methods

Lasso estimation

- Tibshirani (1996): Least absolute shrinkage and selection operator (lasso)

$$\boldsymbol{B}_N^{(L)} = \arg\min_{\boldsymbol{\beta}} \left(\boldsymbol{Y}_U - \boldsymbol{X}_U\boldsymbol{\beta}\right)^T \left(\boldsymbol{Y}_U - \boldsymbol{X}_U\boldsymbol{\beta}\right) + \lambda_N \sum_{j=1}^{p} |\beta_j|$$

  - simultaneously performs model selection and estimation by shrinking unnecessary coefficients to zero

Lasso survey regression estimator

- Construct survey-weighted version:

$$\widehat{\boldsymbol{B}}_N^{(L)} = \arg\min_{\boldsymbol{\beta}} \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}\right) + \lambda_N \sum_{j=1}^{p} |\beta_j|$$

- Plug into model-assisted form:

$$\text{LASSO}(y_k) = \sum_{k\in U} \boldsymbol{x}_k' \widehat{\boldsymbol{B}}_N^{(L)} + \sum_{k\in s} \frac{y_k - \boldsymbol{x}_k' \widehat{\boldsymbol{B}}_N^{(L)}}{\pi_k}$$

- Assume conditions under which GREG is consistent and asymptotically equivalent to the difference estimator,

$$\sum_{k\in U} \boldsymbol{x}_k' \boldsymbol{B}_N + \sum_{k\in s} \frac{y_k - \boldsymbol{x}_k' \boldsymbol{B}_N}{\pi_k}, \text{ where } \boldsymbol{B}_N = \left(\boldsymbol{X}_U^T \boldsymbol{X}_U\right)^{-1} \boldsymbol{X}_U^T \boldsymbol{y}_U$$

- If $\lambda_N = o\left(\sqrt{N}\right)$, then LASSO shares this asymptotic equivalence
  - working model $\neq$ true model, so "oracle properties" are not relevant
  - any advantages of LASSO are in finite samples

- Can also consider adaptive lasso, "A": Zou (2006, *JASA*)
- Lasso is non-linear, so get weights either by
  - Wu and Sitter (2001) model calibration: "C"
  - Tibshirani (1996) ridge regression approximation to lasso coefficient estimates: "R"

nonlinear	calibration weights	ridge weights
LASSO	CLASSO	RLASSO
ALASSO	CALASSO	RALASSO

Study finite sample properties via simulation

- Utah tree canopy cover data set
  - national pilot project conducted by FIA and the US Forest Service Remote Sensing Applications Center
  - $N = 4,151$ grid points on one Landsat scene in Utah
  - covers parts of 10 counties
- Response: $y_k =$ photo-interpreted tree canopy cover
  - relevant to forest management, fire modeling, air pollution mitigation, water temperature, and carbon storage
  - correlated with many other interesting responses
  - known for all $k \in U$ for this pilot study
  - very expensive to obtain!

Utah tree canopy cover simulation, continued

- Auxiliary data $\boldsymbol{x}_k$ from the 2001 National Land Cover Database and Landsat-5 reflectance bands
  - transformed aspect, slope, topographic positional index, elevation, land cover and NLCD predicted tree canopy cover
  - about half are statistically significant in finite population regression of y on $\boldsymbol{x}$
- Sampling designs:
  - simple random (SI) and stratified simple random (STSI)with counties as strata
  - samples sizes $n = 50$ and $n = 100$
  - equal allocation to counties $\Rightarrow$ unequal probabilities
  - 2000 replicate samples from same fixed, finite population

Results: Design bias and MSE

- For all estimators, relative design bias was less than 2%
- Ratio of design MSE for each estimator to design MSE of full GREG estimator:

		SI		**STSI**	
		$n = 50$	$n = 100$	$n = 50$	$n = 100$
Model-Assisted:	LASSO	0.69	0.92	0.78	0.92
	ALASSO	0.76	0.94	0.80	0.94
	CLASSO	0.70	0.91	0.79	0.93
	CALASSO	0.77	0.94	0.83	0.94
	RLASSO	0.69	0.92	0.77	0.92
	RALASSO	0.71	0.97	0.69	0.95
Design Only:	HT	1.29	2.01	1.24	1.73

Results: Weight properties

- Negative weights:
  - calibration weights were negative in only 0.036% of all cases
  - ridge regression weights in only 0.65%
  - GREG weights varied from 1% to 14% negative weights
- Weight variation within samples and across samples:

	Weight Variances							
	$\overline{var}(\boldsymbol{w})$				$\overline{var}(w_j \mid j \in s)$			
	SI		**STSI**		**SI**		**STSI**	
Estimators	$n = 50$	$n = 100$	$n = 50$	$n = 100$	$n = 50$	$n = 100$	$n = 50$	$n = 100$
CLASSO	158	18	5086	1216	155	18	837	92
CALASSO	170	18	5116	1216	166	18	894	93
RLASSO	496	47	6841	1322	441	46	3370	382
RALASSO	236	25	5059	1211	228	25	1059	125
GREG	5261	331	15349	1859	3810	298	12487	1310
HT	0	0	4816	1192	0	0	0	0

Summary

- Lasso survey regression estimators have useful potential
  - dominate MSE of GREG in small samples with large numbers of potential predictors
  - better weight properties: less variation and fewer negative weights
- Sophisticated computational methods can improve basic survey estimators
  - model-assisted framework gives straightforward recipe for incorporating complex methods
- Contact: jbreidt@stat.colostate.edu

Selected references on model-assisted survey regression estimation (by no means exhaustive, sorry!)

- Breidt, F. J., G. Claeskens, and J. D. Opsomer (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika 92*, 831–846.
- Breidt, F. J. and J. D. Opsomer (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics 28*, 1026–1053.
- Breidt, F.J. and Opsomer, J.D. (2008). Endogenous post-stratification in surveys: classifying with a sample-fitted model. *Annals of Statistics 36*, 403–427.
- Breidt, F. J. and J. D. Opsomer (2009). Nonparametric and semiparametric estimation in complex surveys. *Sample Surveys: Theory, Methods and Inference, Handbook of Statistics 29*, 103–119.
- 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.
- Cassel, C. M., C. E. Särndal, and J. H. Wretman (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika 63*, 615–620.
- Dahlke, M., Breidt, F.J., Opsomer, J.D. and Van Keilegom, I. (2013). Nonparametric endogenous post-stratification estimation. *Statistica Sinica 23*, 189-211
- Deville, J.C., and Särndal, C.E. (1992), Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association 87*, 376-382.
- Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliarie: Une approache non paramétrique par splines de régression. *Canadian Journal of Statistics 33*, 163–180.

- McConville, K. S. and F. J. Breidt (2013). Survey design asymptotics for the model-assisted penalised spline regression estimator. *Journal of Nonparametric Statistics* (ahead-of-print), 1–19.
- Montanari, G. and M. Ranalli (2005). Nonparametric methods in survey sampling. *New Developments in Classification and Data Analysis 100*, 203–210.
- Montanari, G. E. and M. G. Ranalli (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association 100(472)*, 1429–1442.
- Montanari,G.E.,and Ranalli, M.G.(2006), A Mixed Model-Assisted Regression Estimator that Uses Variables Employed at the Design Stage, *Statistical Methods and Applications 15*, 139–149.
- Montanari,G.E.,and Ranalli, M.G. (2009).Multiple and ridge model calibration. In *Proc. Workshop on Calibration and Estimation in Surveys.* Statistics Canada.
- Opsomer, J. D., F. J. Breidt, G. G. Moisen, and G. Kauermann (2007). Model-assisted estimation of forest resources with generalized additive models (with discussion). *Journal of the American Statistical Association 102*, 400–416.
- Park, M. and Fuller, W.A. (2005), Towards nonnegative regression weights for survey samples. *Survey Methodology 31*, 85–93.
- Park, M. and Fuller, W.A. (2009), The mixed model for survey regression estimation, *Journal of Statistical Planning and Inference 139*, 1320–1331.
- Rao, J.N.K. and Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling In *ASA Proc. Sect. Survey Res. Meth.*, American Statistical Association.
- Robinson, P. M. and C. E. Särndal (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya: The Indian Journal of Statistics, Series B 45*, 240–248.
- Särndal, C. E. (2007). The calibration approach in survey theory and practice. *Survey Methodology 33(2)*, 99–119.

- Särndal, C.-E., B. Swensson, and J. Wretman (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika 76*, 527–537.
- Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Wang, L. and S. Wang (2011). Nonparametric additive model-assisted estimation for survey data. Journal of Multivariate Analysis 102, 1126–1140.
- Wu, C. and R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association 96*, 185–193.