Nonparametric Endogenous Post-Stratification Estimation

Jay Breidt
Colorado State University
JSM 2010 Vancouver

Joint work with Mark Dahlke and Jean Opsomer, Colorado State University and Ingrid Van Keilegom, Université Catholique de Louvain

Research supported by US National Science Foundation (SES–0922142).
US Forest Service’s Forest Inventory and Analysis (FIA)

- 400K sample plots nationwide, visited every 5 years in rotating panels
- National and regional estimates of forest area, wood volume, growth, mortality, ...
Instead of stratification, FIA uses post-stratification

• Samples the landscape with uniform design, ignoring stratum information

• Incorporates stratum information at the estimation stage

• Counts sample elements falling into post-strata: Random post-stratum sample size

• Reweights sample landscape proportions to agree with known population landscape proportions
Landscape Proportions from FIA Mapping Efforts

- Have extensive spatial information based on satellite imagery (MODIS, TM), topography, other maps, ...
- Active effort within FIA Program to develop predictive maps of forest resources
- Integrate forest data (survey data and otherwise) into GIS for many uses:
  - estimation of damage from wildfires, fire risk
  - delineate areas of harvestable lumber
  - delineate forest/non-forest
Incorporating Landscape-Level Auxiliary Information

- Remotely-sensed image of entire landscape: $\{x_i\}, i \in U_N$.
- LandSat image of 555 km$^2$ Hayman Springs fire scar, Colorado USA
• Constructing post-strata from image data \((\{x_i\}, i \in U_N)\) requires image classification

• FIA survey data \((\{z_i\}, i \in s)\) used as ground truth for “training” image classification algorithms (regression, nonparametric regression, neural nets, classification trees, . . .)
  – Survey data trains classification algorithm
  – Classified image stratifies survey data

• Post-stratification becomes *endogenous* to sample
• Post-stratification is endogenous to the sample

• **Issue:** is it valid to use FIA-based image classification maps as controls?
  – population proportion is now random, not fixed
  – sample proportion is now subject to misclassification error (not just sampling error)

• Both violate classical assumptions for validity of post-stratification
• Introduce modeling framework for classification

• Review results for parametric, generalized linear model: EPSE (Breidt and Opsomer 2008 Ann. Stat.)

• New results for Nonparametric EPSE = NEPSE
  – asymptotic behavior: consistent variance estimator and CLT in model-based setting
  – finite-sample behavior: simulation results with penalized splines in design-based setting
• Have covariate vectors $x_i, i \in U_N$

• Classification index $m(x_i)$ partitions $U_N$ into $H$ post-strata according to boundaries

\[-\infty \leq \tau_0 < \tau_1 < \cdots < \tau_{H-1} < \tau_H \leq \infty\]

• Given $m$, index maps image to post-strata without error

• Parametric case: $m(x_i) = m\lambda(x_i)$; $\lambda$ known implies error-free classification
• Need a classification index $m(\cdot)$ from the image to the classes.
Consider equal-probability sampling.

Define for $\ell = 0, 1, 2$:

$$A_{N hl}(m) = \sum_{i \in U_N} \frac{y_i^{\ell}}{N} I\{\tau_{h-1} < m(x_i) \leq \tau_h\}$$

$$A_{n hl}(m) = \sum_{i \in s_N} \frac{y_i^{\ell}}{n} I\{\tau_{h-1} < m(x_i) \leq \tau_h\}$$
• Estimation target is

\[ \mu_y = \sum_{h=1}^{H} \frac{A_{Nh0}(m)}{A_{Nh0}(m)} A_{Nh1}(m). \]

• Post-stratification estimator (known \( m \)) is

\[ \hat{\mu}(m) = \sum_{h=1}^{H} \frac{A_{Nh0}(m)}{A_{nh0}(m)} A_{nh1}(m). \]
Estimating the Classification Index

- For $m(\cdot)$ unknown, estimate via nonparametric regression of survey data $z_i$ on $x_i$, where $E[z_i | x_i] = m(x_i)$. 

![Diagram of estimating the Classification Index](image)
- Post-stratification estimator (known $m$) is

$$
\hat{\mu}(m) = \sum_{h=1}^{H} \frac{A_{Nh0}(m)}{A_{nh0}(m)} A_{nh1}(m).
$$

- Endogenous post-stratification estimator is

$$
\hat{\mu}(\hat{m}) = \sum_{h=1}^{H} \frac{A_{Nh0}(\hat{m})}{A_{nh0}(\hat{m})} A_{nh1}(\hat{m}).
$$

- In particular, parametric EPSE is

$$
\hat{\mu} \left( m_{\hat{\lambda}} \right) = \sum_{h=1}^{H} \frac{A_{Nh0} \left( m_{\hat{\lambda}} \right)}{A_{nh0} \left( m_{\hat{\lambda}} \right)} A_{nh1} \left( m_{\hat{\lambda}} \right).
$$
Studying Properties of EPSE

\[ \hat{\mu}(\hat{m}) = \sum_{h=1}^{H} \frac{A_{Nh0}(\hat{m})}{A_{nh0}(\hat{m})} A_{nh1}(\hat{m}). \]

- Intuitively, if \( \hat{m}(\cdot) \) is a good estimator of \( m(\cdot) \), EPSE should be close to PSE and share its statistical properties.
- But EPSE is complicated nonlinear and non-differentiable function of sample quantities.
Parametric EPSE Properties

- Design-consistent under mild conditions.
- Central limit theorem:
  \[
  \left\{ \frac{1}{n} \left( 1 - \frac{n}{N} \right) \hat{V}_y(m_\lambda) \right\}^{-1/2} (\hat{\mu}(m_\lambda) - \mu_y) \xrightarrow{\mathcal{L}} N(0, 1)
  \]
  as \( n, N \to \infty \), where
  \[
  \hat{V}_y(m_\lambda) = \sum_{h=1}^{H} \frac{A_{Nh0}^2(m_\lambda)}{A_{nh0}(m_\lambda)} \frac{A_{nh2}(m_\lambda) - A_{nh1}^2(m_\lambda)}{A_{nh0}(m_\lambda) - n^{-1}}
  \]
  is consistent for
  \[
  V_y(m_\lambda) = \sum_{h=1}^{H} \text{Pr} \left[ \tau_{h-1} < m_\lambda(x_i) \leq \tau_h \right] \text{Var} \left( y_i \mid \tau_{h-1} < m_\lambda(x_i) \leq \tau_h \right)
  \]
- Same as PSE results, modulo estimation of \( \lambda \) by \( \hat{\lambda} \)
Proof for Parametric EPSE

• Taylor linearization? \( \hat{\mu}(m^{\hat{\lambda}}) - \hat{\mu}(m_\lambda) = o_P \left( n^{-1/2} \right) \)

• Technical problem is presence of \( \hat{\lambda} \) in non-differentiable indicators: \( I_{\{\tau_{h-1} < m^{\hat{\lambda}}(x_i) \leq \tau_h}\} \)

• Approach: \( A_{Nh\ell}(m_\lambda), A_{nh\ell}(m_\lambda) \) are U-statistics with kernel \( y_i \ell I_{\{\tau_{h-1} < m_\lambda(x_i) \leq \tau_h\}} \) and expectation \( \alpha_{h\ell}(m_\lambda) \)

• By verifying conditions in Randles (1982, *Ann. Statist.*), we establish appropriate linearization results:

\[
A_{Nh\ell}(m^{\hat{\lambda}}) - \alpha_{h\ell}(m^{\hat{\lambda}}) - A_{Nh\ell}(m_\lambda) + \alpha_{h\ell}(m_\lambda) = o_P \left( N^{-1/2} \right)
\]

\[
A_{nh\ell}(m^{\hat{\lambda}}) - \alpha_{h\ell}(m^{\hat{\lambda}}) - A_{nh\ell}(m_\lambda) + \alpha_{h\ell}(m_\lambda) = o_P \left( n^{-1/2} \right)
\]
• Have the same technical problem, but worse: presence of \( \hat{m} \) in non-differentiable indicators: \( I_{\{\tau_{h-1} < \hat{m}(x_i) \leq \tau_h\}} \)

• Requires different technical approach to establish the following “linearization” lemma:

\[
A_{Nh\ell}(\hat{m}) - \alpha_{h\ell}(\hat{m}) - A_{Nh\ell}(m) + \alpha_{h\ell}(m) = o_P\left(N^{-1/2}\right)
\]

\[
A_{nh\ell}(\hat{m}) - \alpha_{h\ell}(\hat{m}) - A_{nh\ell}(m) + \alpha_{h\ell}(m) = o_P\left(n^{-1/2}\right)
\]

• Use results from empirical process theory: want \( m \) to live in a Donsker class
Asymptotics for NEPSE in equal-probability, model-based setting rely on fairly standard model assumptions:

• A1. \( \{x_i\} \) iid, non-degenerate continuous random vectors with compact support; \( \Pr [m(x) \leq u] \) is Lipschitz continuous in \( u \) of order \( 0 < \alpha \leq 1 \).

• A2. Study variables \( y | x \) are conditionally independent with \( \mathbb{E} \left[ y^4 | x \right] < \infty \) and

\[
\alpha_{\tau\ell}(m) = \mathbb{E} \left[ y_i^\ell I\{m(x_i) \leq \tau\} \right]
\]

continuous in \( m \) for \( \ell = 0, 1, 2 \), and \( \alpha_{\tau h 0}(m) > \alpha_{\tau h -1 0}(m) \)
Asymptotics for NEPSE also rely on method assumptions, using results from empirical process theory.

- **A3.** Nonparametric estimator $\hat{m}(\cdot)$ satisfies
  \[
  \sup_\mathbf{x} |\hat{m}(\mathbf{x}) - m(\mathbf{x})| = o(1) \text{ a.s.}
  \]

- **A4.** There exists a space $\mathcal{D}$ of measurable functions that satisfies $m \in \mathcal{D}$, $\Pr[\hat{m} \in \mathcal{D}] \to 1$ as $n \to \infty$, and
  \[
  \int_0^\infty \sqrt{\log N[\cdot]}(\lambda, \mathcal{F}, \| \cdot \|_2) \, d\lambda < \infty
  \]
  where $N[\cdot]$ is the bracketing number and
  \[
  \mathcal{F} = \{ \mathbf{x} \to I(d(\mathbf{x}) \leq \tau) : d \in \mathcal{D} \}.
  \]
Verifying A4: Some Examples

- **Monotone**: \( m \in \mathcal{D} = \) monotone, bounded functions over a compact subset of \( \mathbb{R} \)

- **Partially linear monotone**: \( \mathbf{x} = (x_1, x_2), m \in \mathcal{D} = \) functions of the form \( \beta' x_1 + d(x_2), d \) monotone as above.

- **Single index monotone**: \( m \in \mathcal{D} = \) functions of the form \( d (\beta' \mathbf{x}), d \) monotone as above.

- Note that the *model* has a monotone component, but the *method* does not require monotonicity: classical local polynomials or penalized splines work.
Nonparametric EPSE Properties

- Central limit theorem:

\[
\left\{ \frac{1}{n} \left(1 - \frac{n}{N}\right) \hat{V}_y (\hat{m}) \right\}^{-1/2} (\hat{\mu} (\hat{m}) - \mu_y) \xrightarrow{L} N(0, 1)
\]

as \( n, N \to \infty \), where

\[
\hat{V}_y (\hat{m}) = \sum_{h=1}^{H} \frac{A_{N h 0}^2 (\hat{m})}{A_{n h 0} (\hat{m})} \frac{A_{n h 2} (\hat{m}) - A_{n h 1}^2 (\hat{m}) A_{n h 0}^{-1} (\hat{m})}{A_{n h 0} (\hat{m}) - n^{-1}}
\]

is consistent for

\[
V_y (m) = \sum_{h=1}^{H} \Pr [\tau_{h-1} < m(x_i) \leq \tau_h] \Var (y_i \mid \tau_{h-1} < m(x_i) \leq \tau_h)
\]

- Same as PSE results, modulo estimation of \( m \) by \( \hat{m} \)
Simulation Setup

- Population with $N = 1000$
- Variable $x_i \sim \text{Unif}(0, 1)$, variable $z_i$ linear in $x_i$
- 1000 replications of simple random sampling
- Estimators for population means
  - sample mean (HTE = Horvitz-Thompson estimator)
  - survey regression estimator (REG)
  - PSE on $m(x_i)$ with 4 strata
  - endogenous PSE with 4 strata: penalized spline of $z_i$ on $x_i$ with df $= 2$ (EPSE) or df $= 5$ (NEPSE)
• Image variable $x_i \sim \text{Unif}(0, 1)$, training variable $z_i$ linear in $x_i$, seven additional survey variables
Simulation Results for $n = 50$

- Ratio of estimator MSE to EPSE ($df = 2$) MSE, NEPSE ($df = 5$) MSE

<table>
<thead>
<tr>
<th>Response</th>
<th>HTE</th>
<th>PSE</th>
<th>REG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line</td>
<td>4.98, 4.68</td>
<td>1.10, 0.95</td>
<td>0.74, 0.69</td>
</tr>
<tr>
<td>Quad</td>
<td>2.34, 2.29</td>
<td>1.03, 1.01</td>
<td>2.56, 2.51</td>
</tr>
<tr>
<td>Bump</td>
<td>3.22, 3.26</td>
<td>1.00, 1.01</td>
<td>0.94, 0.95</td>
</tr>
<tr>
<td>Jump</td>
<td>2.19, 2.13</td>
<td>1.00, 0.97</td>
<td>1.80, 1.76</td>
</tr>
<tr>
<td>Curve</td>
<td>1.88, 1.88</td>
<td>0.99, 0.99</td>
<td>1.17, 1.17</td>
</tr>
<tr>
<td>Cycle1</td>
<td>3.10, 3.04</td>
<td>1.04, 1.02</td>
<td>1.56, 1.53</td>
</tr>
<tr>
<td>Cycle4</td>
<td>0.96, 0.98</td>
<td>1.00, 1.02</td>
<td>0.92, 0.94</td>
</tr>
<tr>
<td>Noise</td>
<td>0.93, 0.92</td>
<td>1.00, 0.99</td>
<td>0.96, 0.95</td>
</tr>
</tbody>
</table>
Findings: Endogenous Post-Stratification is OK

- Analytical summary: NEPSE, EPSE are asymptotically equivalent to usual PSE
- Simulation summary:
  - NEPSE (df = 5) and EPSE (df = 2) are nearly identical to PS estimator even for small sample sizes ($n = 50$)
  - NEPSE, EPSE generally improve over HTE
  - Some loss of efficiency compared to REG when mean function properly specified
  - Not shown: variance estimator has some negative bias, but 95% confidence intervals based on CLT have close to nominal coverage (93%–96%)