Nonparametric Model-Assisted Estimation of Distribution Functions from Survey Data

F. Jay Breidt
Colorado State University

Joint work with Alicia A. Johnson, Colorado State University and Jean D. Opsomer, Iowa State University

The work reported here was developed under STAR Research Assistance Agreements CR-829095 and CR-829096 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University and Oregon State University. This presentation has not been formally reviewed by EPA. The views expressed here are solely those of the authors. EPA does not endorse any products or commercial services mentioned in this report.
Specific versus Generic

- **Specific (brand-name?):** not a black box; expensive; very good for its purpose
- **Generic:** pretty much a black box; cheap; good for many purposes
Finite Population CDF Estimation

• Some notation:

\[ F(t) = \frac{1}{N} \sum_{i \in U} I\{y_i \leq t\} \]

where \( U = \{1, 2, \ldots, N\} \) ("landscape")

\(- y_i \) observed for sample \( s \subset U \) of size \( n \)

\(- \) auxiliary information \( x_i \) available for all of \( U \)

• Idea: Use auxiliary information to improve finite population distribution function estimation

\(- \) model \( (x_i, y_i) \) relationship and use to predict non-sampled \( y_i, i \in U - s \)
Modeling Contexts

• **Specific:** few study variables, few population parameters
  
  – lots of modeling resources to specify, estimate, and diagnose a model
  
  – willingness to defend the model

• **Generic:** many study variables, many population parameters
  
  – no resources to model every variable
  
  – no single model is adequate/defensible
Generic Inferences in Natural Resources Monitoring

- **Example:** conduct a survey and prepare a report
  - analyze large numbers of chemical, biological, and physical variables
  - estimate means, quantiles, and distribution functions
  - break down both by political classifications and by various ecological classifications

- Generic inference is a common problem for federal, state, and tribal agencies
Very Generic: Horvitz-Thompson Estimator

- Let $\pi_i = \Pr \{i \in s\}$, $\pi_{ij} = \Pr \{i, j \in s\}$ and $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$
- Then

$$\hat{F}_{HT}(t) = \frac{1}{N} \sum_{i \in s} \frac{I\{y_i \leq t\}}{\pi_i}$$

and

$$\operatorname{Var}(\hat{F}_{HT}(t)) = \frac{1}{N^2} \sum_{i,j \in s} \frac{\Delta_{ij} I\{y_i \leq t\} I\{y_j \leq t\}}{\pi_{ij} \pi_i \pi_j}$$

are design unbiased and consistent
- no dependence on any model
- does not incorporate auxiliary information $x_i$
Estimation with Auxiliary Information

- **Superpopulation model:**
  \[ y_i = m(x_i) + v^{1/2}(x_i)\epsilon_i \]
  where \( \epsilon_i \sim G \) with \( E(\epsilon_i) = 0, \text{Var}(\epsilon_i) = \sigma^2 \)

- **Model-based:** biased if model is wrong
  \[ \sum_{i \in U - s} \text{(model-based prediction)} + \sum_{i \in s} \text{(sampled values)} \]

- **Model-assisted:** design-unbiased even if model is wrong
  \[ \sum_{i \in U} \text{(model-based prediction)} + \text{(design bias adjustment)} \]
Model-Based Parametric: CD Estimator

- Chambers and Dunstan (1986)
  - model-based

\[
\hat{F}_{CD}(t) = \frac{1}{N} \sum_{i \in U - s} \hat{G}_i + \frac{1}{N} \sum_{i \in s} I\{y_i \leq t\}
\]

- \(\hat{G}_i\) estimates \(G\left(\frac{t - m(x_i)}{v^{1/2}(x_i)}\right) = E_m\left(I\{y_i \leq t\}\right)\)
- asymptotically unbiased when \(m(x_i)\) and \(v(x_i)\) correctly specified
Model-Assisted Parametric: RKM Estimator

- Rao, Kovar, Mantel (1990)
  - model-assisted

\[
\hat{F}_{RKM}(t) = \frac{1}{N} \sum_{i \in U} \tilde{G}_i + \sum_{i \in s} \frac{I\{y_i \leq t\} - \tilde{G}_{ic}}{N\pi_i}
\]

where \(\tilde{G}_{ic}\) is \(\tilde{G}_i\) weighted with conditional probabilities

- asymptotically design and model unbiased
Motivation for Nonparametric Methods

- $E_m I_{y_i \leq t} = \Pr \{ y_i \leq t \} = G(v^{-1/2}(x_i)(t - m(x_i)))$
- Mean function misspecification
  - CD will be biased
  - RKM will be inefficient
  - nonparametric methods only assume $m(x_i)$ is smooth
- Variance misspecification
  - CD and RKM assume $v(x_i)$ is known
  - nonparametric assumes $v(x_i)$ smooth, positive
Possible Smoothing Strategies

- **Specific:** use original response
  - smooth $y_i$ versus $x_i$ to get $\hat{m}(x_i)$
  - smooth $(y_i - \hat{m}(x_i))^2$ versus $x_i$ to get $\hat{v}(x_i)$
  - plug in to CD- or RKM-like estimator

- **Generic:** use indicators
  - smooth $I\{y_i \leq t\}$ versus $x_i$ to get $\hat{G}(v^{-1/2}(x_i)(t - m(x_i)))$
  - plug in to model-based or model-assisted estimator
Possible Smoothing Strategies

- Smooth response $y_i$ or indicator $I_{y_i \leq t}$ versus $x_i$
A Nonparametric Method: Local Polynomial Regression

- Smooth function locally approximated by $q$th-order polynomial
- Sample design matrix $(n \times (q + 1))$:

\[ X_{si} = \begin{bmatrix} 1 & x_j - x_i & \cdots & (x_j - x_i)^q \end{bmatrix}_{j \in s} \]

- Sample weighting matrix $(n \times n)$:

\[ W_{si} = \text{diag} \left\{ \frac{1}{\pi_j h} K \left( \frac{x_j - x_i}{h} \right) \right\}_{j \in s} \]

- Sample smoother vector at $x_i$:

\[ s'_{si} = [1 \ 0 \ \cdots \ 0] \left( X'_{si} W_{si} X_{si} \right)^{-1} X'_{si} W_{si} \]
LPR in Survey Sampling Estimation

- Finite population total: $T_y = \sum_{i \in U} y_i$
- Breidt and Opsomer (2000):

$$\hat{T}_{LPR} = \sum_{i \in U} \hat{m}_i + \sum_{i \in s} \frac{y_i - \hat{m}_i}{\pi_i}$$

where $\hat{m}_i = s'_{si}[y_i]_{i \in s}$, and

$$\text{Var}(\hat{T}_{LPR}) = \sum_{i, j \in s} \frac{\Delta_{ij}(y_i - \hat{m}_i)(y_j - \hat{m}_j)}{\pi_{ij}\pi_i\pi_j}$$

are asymptotically design unbiased and consistent
- comparable efficiency to REG under linear model
- more efficient than REG otherwise
Local Polynomial Regression CDF Estimator

- Model-assisted approach
- Define $I_s = [I_{y_i \leq t}]_{i \in s}$
- Estimate $\hat{G}(v^{-1/2}(x_i)(t - m(x_i)))$ by $\hat{g}_i = s_{si}^t I_s$
- Then
  $$\hat{F}_{LPR}(t) = \frac{1}{N} \sum_{i \in U} \hat{g}_i + \frac{1}{N} \sum_{i \in s} \frac{I_{y_i \leq t}}{\pi_i} - \hat{g}_i$$
- Can construct model-based nonparametric estimator analogously
Properties of LPR CDF Estimator

• From Breidt and Opsomer (2000), $\hat{F}_{LPR}(t)$ and

$$\text{Var}(\hat{F}_{LPR}(t)) = \frac{1}{N^2} \sum_{i,j \in s} \frac{\Delta_{ij}(I\{y_i \leq t\} - \hat{g}_i)(I\{y_j \leq t\} - \hat{g}_j)}{\pi_i \pi_j}$$

are asymptotically design unbiased and consistent

• Weighted form for generic inference:

$$\hat{F}_{LPR}(t) = \sum_{i \in s} \omega_{is} I\{y_i \leq t\},$$

where the weights $\{\omega_{is}\}$ do not depend on $y$ or $t$ – can be applied to any response at any quantile
Internal Consistency

• LPR weights guarantee internal consistency:
  \[
  \sum_{i \in s} \omega_i s (y_i + z_i) = \sum_{i \in s} \omega_i s y_i + \sum_{i \in s} \omega_i s z_i
  \]

• Mean of LPR-estimated cdf is LPR-estimated mean
  \[
  \int y \, d\hat{F}_{LPR}(y) = \sum_{i \in s} \omega_i s y_i = \frac{T_{LPR}}{N}
  \]
  – assuming weights are non-negative
Range of Estimators

- From specific to generic:

<table>
<thead>
<tr>
<th>Specific</th>
<th>Model-based parametric</th>
<th>CD</th>
</tr>
</thead>
<tbody>
<tr>
<td>↓</td>
<td>Model-assisted parametric</td>
<td>RKM</td>
</tr>
<tr>
<td>↓</td>
<td>Model-based nonparametric</td>
<td>DORF</td>
</tr>
<tr>
<td>↓</td>
<td>Model-assisted nonparametric</td>
<td>LPR</td>
</tr>
<tr>
<td>Generic</td>
<td>Design-based</td>
<td>HT</td>
</tr>
</tbody>
</table>
CDF Simulation Study Design

- Compare seven estimators via simulation:

<table>
<thead>
<tr>
<th>Type</th>
<th>Estimator</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>design-based</td>
<td>HT</td>
<td>no model</td>
</tr>
<tr>
<td>model-based</td>
<td>CD0</td>
<td>$\beta_1 x + x^{1/2}\epsilon$</td>
</tr>
<tr>
<td></td>
<td>CD1</td>
<td>$\beta_0 + \beta_1 x + \epsilon$</td>
</tr>
<tr>
<td></td>
<td>DORF</td>
<td>$m(x) + v^{1/2}(x)\epsilon$</td>
</tr>
<tr>
<td>model-assisted</td>
<td>RKM0</td>
<td>$\beta_1 x + x^{1/2}\epsilon$</td>
</tr>
<tr>
<td></td>
<td>RKM1</td>
<td>$\beta_0 + \beta_1 x + \epsilon$</td>
</tr>
<tr>
<td></td>
<td>LPR</td>
<td>$m(x) + v^{1/2}(x)\epsilon$</td>
</tr>
</tbody>
</table>
Simulated Response Variables

- 7 response variables with $x_i \sim \text{Unif}(0, 1)$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
CDF Simulation Study Design, Continued

- \(N = 1000, n = 100, \pi_i = n/N\)
- 1000 reps
- \(\hat{F}_{LPR}\) calculated using Epanechnikov kernel:
  \[
  K(x) = \frac{3}{4}(1 - x^2)I\{|x|<1\}
  \]
- Bandwidth \(h = 0.1\) or \(0.25\)
  - single choice of \(h\) is not optimal
  - single choice means one generic set of weights, \(\{\omega_{is}\}\)
- \(\sigma = 0.1\) or \(0.4\)
- CDF estimated at median and first quartile
CDF Simulation Study Output

- Return MSE ratios: (> 1 favors LPR)
  \[
  \frac{MSE(\hat{F}_{\ast}(t))}{MSE(\hat{F}_{LPR}(t))}
  \]

- Return percent relative biases:
  \[
  \left(\frac{\hat{F}(t) - F(t)}{F(t)}\right) 100\%
  \]
Smoothing to Estimate the Linear CDF

- Linear mean versus $x$
- Linear CDF
- Indicator mean at quartile vs. $x$
- Indicator mean at median vs. $x$
Smoothing to Estimate the Bump CDF
Smoothing to Estimate the Jump CDF

Jump mean versus $x$

Jump CDF

Indicator mean at quartile vs. $x$

Indicator mean at median vs. $x$
CDF Simulation Results: Bias

- \((2 \sigma)(2 \text{ bandwidths})(7 \text{ responses})(2 \text{ quantiles}) = 56 \text{ cases}\)

\[\text{RB} = \text{Relative Bias}\]

<table>
<thead>
<tr>
<th>Design-based</th>
<th>HT</th>
<th>&lt; 0.5% RB (\leq 2%) RB (\geq 2%) RB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model-assisted</td>
<td>RKM0, RKM1, LPR</td>
<td>(\leq 2%) RB (\geq 2%) RB (\geq 2%) RB</td>
</tr>
<tr>
<td>Model-based par.</td>
<td>CD0, CD1</td>
<td>(\geq 2%) RB (\geq 2%) RB (\geq 2%) RB</td>
</tr>
<tr>
<td>Model-based nonpar.</td>
<td>DORF</td>
<td>(\geq 2%) RB (\geq 2%) RB (\geq 2%) RB</td>
</tr>
</tbody>
</table>

- Design bias adjustment works!
CDF Simulation Numerical Results

- MSE ratios for CDF estimation at the median, $h = 0.25$, $\sigma = 0.4$

<table>
<thead>
<tr>
<th>Response</th>
<th>HT</th>
<th>CD0</th>
<th>CD1</th>
<th>RKM0</th>
<th>RKM1</th>
<th>DORF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio</td>
<td>1.24</td>
<td>0.71</td>
<td>1.94</td>
<td>0.95</td>
<td>0.97</td>
<td>1.22</td>
</tr>
<tr>
<td>Linear</td>
<td>2.16</td>
<td>2.86</td>
<td>0.56</td>
<td>0.97</td>
<td>0.97</td>
<td>1.40</td>
</tr>
<tr>
<td>Expo</td>
<td>1.06</td>
<td>1.02</td>
<td>0.83</td>
<td>1.20</td>
<td>0.99</td>
<td>1.17</td>
</tr>
<tr>
<td>Bump</td>
<td>2.26</td>
<td>6.36</td>
<td>2.62</td>
<td>1.08</td>
<td>1.14</td>
<td>1.39</td>
</tr>
<tr>
<td>Jump</td>
<td>1.26</td>
<td>1.26</td>
<td>0.95</td>
<td>1.13</td>
<td>1.18</td>
<td>1.24</td>
</tr>
<tr>
<td>Quad</td>
<td>1.04</td>
<td>0.51</td>
<td>0.97</td>
<td>1.34</td>
<td>1.05</td>
<td>1.20</td>
</tr>
<tr>
<td>Cycle</td>
<td>2.79</td>
<td>3.11</td>
<td>1.19</td>
<td>4.29</td>
<td>1.38</td>
<td>1.57</td>
</tr>
</tbody>
</table>

- $m(x)$ not misspecified
- $m(x)$ misspecified
MSE Comparisons to Generic Estimators

- LPR dominates HT and DORF in nearly all cases

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Q1</th>
<th>Med</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT</td>
<td>0.91</td>
<td>1.22</td>
<td>2.01</td>
<td>3.16</td>
<td>10.25</td>
</tr>
<tr>
<td>DORF</td>
<td>0.98</td>
<td>1.13</td>
<td>1.38</td>
<td>1.63</td>
<td>3.29</td>
</tr>
</tbody>
</table>

- LPR is competitive with RKM0, RKM1

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Q1</th>
<th>Med</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>RKM0</td>
<td>0.69</td>
<td>0.94</td>
<td>1.16</td>
<td>1.85</td>
<td>16.55</td>
</tr>
<tr>
<td>RKM1</td>
<td>0.69</td>
<td>0.96</td>
<td>1.03</td>
<td>1.37</td>
<td>4.65</td>
</tr>
</tbody>
</table>
### MSE Comparisons for Specific Estimators

#### Mean correct, variance correct

<table>
<thead>
<tr>
<th></th>
<th>$\sigma = 0.1$</th>
<th>$\sigma = 0.4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD0</td>
<td>0.14, 0.14, 0.39, 0.39</td>
<td>0.56, 0.60, 0.67, 0.71</td>
</tr>
<tr>
<td>CD1</td>
<td>0.17, 0.19, 0.19, 0.23</td>
<td>0.54, 0.56, 0.58, 0.60</td>
</tr>
</tbody>
</table>

#### Mean correct, variance incorrect

<table>
<thead>
<tr>
<th></th>
<th>$\sigma = 0.1$</th>
<th>$\sigma = 0.4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD0</td>
<td>0.07, 0.08, 0.55, 0.67</td>
<td>1.15, 1.21, 2.74, 2.86</td>
</tr>
<tr>
<td>CD1</td>
<td>0.27, 0.27, 0.74, 0.75</td>
<td>0.71, 0.77, 1.85, 1.94</td>
</tr>
</tbody>
</table>

#### Mean incorrect

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Q1</th>
<th>Med</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD0</td>
<td>0.48</td>
<td>1.01</td>
<td>2.71</td>
<td>14.98</td>
<td>84.39</td>
</tr>
<tr>
<td>CD1</td>
<td>0.64</td>
<td>0.91</td>
<td>1.44</td>
<td>3.68</td>
<td>18.36</td>
</tr>
</tbody>
</table>
Quantile Estimation Simulation

- Quantile is \( \theta(\alpha) = \min\{t : F(t) \geq \alpha\} \)
- Estimate by \( \hat{\theta}(\alpha) = \min\{t : \hat{F}(t) \geq \alpha\} \)
- Simulation study design identical to CDF simulation
Results for Estimation of Median

- MSE ratios for median estimation, $h = 0.25$, $\sigma = 0.4$

<table>
<thead>
<tr>
<th>Population</th>
<th>HT</th>
<th>CD0</th>
<th>CD1</th>
<th>RKM0</th>
<th>RKM1</th>
<th>DORF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio</td>
<td>1.26</td>
<td>0.64</td>
<td>1.90</td>
<td>0.97</td>
<td>0.99</td>
<td>1.03</td>
</tr>
<tr>
<td>Linear</td>
<td>2.57</td>
<td>3.77</td>
<td>0.61</td>
<td>1.08</td>
<td>1.08</td>
<td>1.18</td>
</tr>
<tr>
<td>Expo</td>
<td>1.06</td>
<td>0.94</td>
<td>0.97</td>
<td>1.21</td>
<td>1.01</td>
<td>1.02</td>
</tr>
<tr>
<td>Bump</td>
<td>2.37</td>
<td>6.22</td>
<td>1.99</td>
<td>1.12</td>
<td>1.17</td>
<td>1.16</td>
</tr>
<tr>
<td>Jump</td>
<td>1.26</td>
<td>1.85</td>
<td>0.88</td>
<td>1.14</td>
<td>1.18</td>
<td>1.07</td>
</tr>
<tr>
<td>Quad</td>
<td>1.02</td>
<td>2.71</td>
<td>0.92</td>
<td>1.33</td>
<td>1.02</td>
<td>1.02</td>
</tr>
<tr>
<td>Cycle</td>
<td>3.52</td>
<td>16.68</td>
<td>0.78</td>
<td>5.51</td>
<td>1.51</td>
<td>1.55</td>
</tr>
</tbody>
</table>

- Results very similar to CDF simulation results for estimation at the median
Summary

• Finite population CDF estimation
  – incorporation of auxiliary information
  – comparison of generic vs. specific inference

• In generic context, nonparametric model-based (LPR)
  – dominates design-based (HT) and model-based nonparametric (DORF)
  – is competitive with model-assisted parametric (RKM)
  – loses to model-based parametric (CD) for correct mean, correct variance
  – beats CD for incorrect mean or “noticeably” incorrect variance

• Similar results for quantile estimation