Estimating Distribution Functions from Survey Data Using Nonparametric Regression

Alicia Johnson and F. Jay Breidt

Department of Statistics
Colorado State University

Joint work with Jean Opsomer,
Iowa State University

Research Supported by EPA grants:
R-82909501-0 to Colorado State University and
R-82909601-0 to Oregon State University
Outline

• Introduction
  – finite population cdf estimation for $Y$
  – Horvitz-Thompson estimator

• Estimation with auxiliary information
  – auxiliary information $x$ available for entire landscape
  – parametric and nonparametric models, relating $Y$ to $x$
  – motivation for nonparametric methods

• Numerical results
  – Monte Carlo comparison of several estimators
  – mean model misspecification

• Further work
Finite Population CDF Estimation

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

- Some Notation:
  - finite population: \( U = \{1, 2, \ldots, N\} \)
  - \( Y_i \) observed for sample: \( s \subset U \) of size \( n \)
  - \( \pi_i = \Pr \{i \in s\} \)
Horvitz-Thompson Estimator

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

- design unbiased
- no dependence on any model
- does not incorporate auxiliary information \( x \)
- How do we incorporate \( x \) for the entire landscape?
Estimation with Auxiliary Information

<table>
<thead>
<tr>
<th>Model Based</th>
<th>Parametric</th>
<th>Nonparametric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chambers and</td>
<td></td>
<td>Dorfman</td>
</tr>
<tr>
<td>Dunstan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model Assisted</td>
<td>Rao, Kovar, Mantel</td>
<td>LPR</td>
</tr>
</tbody>
</table>

- Model:

\[ Y_i = m(x_i) + u^{1/2}(x_i) \epsilon_i \]

where:

\[ \epsilon_i \sim G \text{ with } E(\epsilon_i) = 0, \text{ Var}(\epsilon_i) = \sigma^2 \]

- \( x_i \) known for all \( i \in U \)
Parametric Methods

\[ Y_i = \beta_0 + \beta_1 x_i + v^{1/2}(x_i) \epsilon_i \]

- \( v^{1/2}(x_i) \) is known and strictly positive
- assumes linear mean function

- CD estimator
  - Chambers and Dunstan (1986)
  - model based

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

where \( \hat{G}_i \) estimates \( G \left( \frac{t - m(x_i)}{v^{1/2}(x_i)} \right) = E_m I\{Y_i \leq t\} \)
Parametric Methods Continued

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

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

\textit{model-based prediction} \quad \textit{design-bias adjustment}

where \( \hat{G}_{ic} \) is \( \hat{G}_i \) weighted with conditional probabilities
Motivation for Nonparametric Methods

Recall: $Y_i = m(x_i) + v^{1/2}(x_i)\epsilon_i$

- mean function misspecification bias
  - CD and RKM assume $m(x_i) = \beta_0 + \beta_1 x_i$ and $v^{1/2}(x_i)$ known
  - if $m(x_i)$ is misspecified:
    * CD will be biased
    * RKM will be inefficient
  - nonparametric methods only assume $m(x_i)$ is smooth

- variance misspecification bias
  - CD and RKM assume $v^{1/2}(x_i)$ is known
Local Polynomial Regression Estimator

- nonparametric, model-assisted
- based on LPR estimator for population total (Breidt and Opsomer, 2000)

\[ \hat{t}_{LPR} = \sum_{i \in U} \hat{m}_i + \sum_{i \in s} \frac{Y_i - \hat{m}_i}{\pi_i} \]

- replace \( Y_i \) with \( I_{\{Y_i \leq t\}} \):

\[ \hat{F}_{LPR}(t) = \sum_{i \in U} \hat{\mu}_i + \sum_{i \in s} \frac{I_{\{Y_i \leq t\}} - \hat{\mu}_i}{\pi_i} \]

where \( \mu \) is a new smooth function of \( x_i \):

\[ E_m(I_{\{Y_i \leq t\}}) = \mu(x_i) = G\left(\frac{t - m(x_i)}{v^{1/2}(x_i)}\right) \]
Study Design

- 7 populations generated from $x_i \sim \text{Unif}(0, 1)$
- Estimators:
  - HT
  - CD0 (no intercept term)
  - CD1
  - RKM0 (no intercept term)
  - RKM1
  - LPR
- simple random sampling ($\pi_i = \frac{n}{N}$)
- model misspecification
Preliminary Numerical Results

- $N = 1000$, $n = 100$, $\sigma = 0.05$
- 100 reps
- return MSE ratios: ($> 1$ favors LPR)

$$\frac{MSE(\hat{F}_*(t))}{MSE(\hat{F}_{LPR}(t))}$$

- CDF estimation at the median

<table>
<thead>
<tr>
<th></th>
<th>ratio1</th>
<th>linear1</th>
<th>expo</th>
<th>bump</th>
<th>jump</th>
<th>quad</th>
<th>cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>hteff</td>
<td>5.45</td>
<td>3.86</td>
<td>1.24</td>
<td>9.11</td>
<td>9.11</td>
<td>1.54</td>
<td>9.93</td>
</tr>
<tr>
<td>cd0eff</td>
<td>0.21</td>
<td>1.60</td>
<td>2.18</td>
<td>19.58</td>
<td>0.79</td>
<td>1.11</td>
<td>15.25</td>
</tr>
<tr>
<td>cd1eff</td>
<td>0.11</td>
<td>0.36</td>
<td>8.75</td>
<td>19.45</td>
<td>3.41</td>
<td>3.11</td>
<td>4.15</td>
</tr>
<tr>
<td>rkm0eff</td>
<td>0.93</td>
<td>1.68</td>
<td>2.54</td>
<td>2.30</td>
<td>2.15</td>
<td>4.51</td>
<td>17.96</td>
</tr>
<tr>
<td>rkm1eff</td>
<td>0.95</td>
<td>0.89</td>
<td>0.96</td>
<td>2.44</td>
<td>2.65</td>
<td>3.60</td>
<td>3.49</td>
</tr>
</tbody>
</table>

NOTE:
m(x) not misspecified

m(x) misspecified
Summary and Further Work

<table>
<thead>
<tr>
<th></th>
<th>Parametric</th>
<th>Nonparametric</th>
</tr>
</thead>
<tbody>
<tr>
<td>model based</td>
<td>Chambers and Dunstan</td>
<td>Dorfman</td>
</tr>
<tr>
<td>model assisted</td>
<td>Rao, Kovar, Mantel</td>
<td>LPR</td>
</tr>
</tbody>
</table>

- variance misspecification bias
- quantile estimation
- analytical comparisons
  (Chambers, Dorfman, Hall (1992))
## Percent Relative Bias Results

<table>
<thead>
<tr>
<th></th>
<th>ratio1</th>
<th>linear1</th>
<th>expo</th>
<th>bump</th>
<th>jump</th>
<th>quad</th>
<th>cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>cd0</td>
<td>1.29</td>
<td>-3.50</td>
<td>6.63</td>
<td>19.55</td>
<td>1.82</td>
<td>-0.96</td>
<td>7.37</td>
</tr>
<tr>
<td>cd1</td>
<td>0.31</td>
<td>1.98</td>
<td>-20.85</td>
<td>19.47</td>
<td>7.94</td>
<td>-0.20</td>
<td>3.29</td>
</tr>
<tr>
<td>rkm0</td>
<td>-0.41</td>
<td>-0.64</td>
<td>0.08</td>
<td>0.39</td>
<td>0.65</td>
<td>0.22</td>
<td>0.26</td>
</tr>
<tr>
<td>rkm1</td>
<td>-0.44</td>
<td>-0.13</td>
<td>-0.24</td>
<td>0.34</td>
<td>0.36</td>
<td>0.72</td>
<td>0.45</td>
</tr>
<tr>
<td>lpr</td>
<td>-0.52</td>
<td>-0.23</td>
<td>-0.02</td>
<td>0.12</td>
<td>-0.04</td>
<td>0.50</td>
<td>0.24</td>
</tr>
</tbody>
</table>
References


