

**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

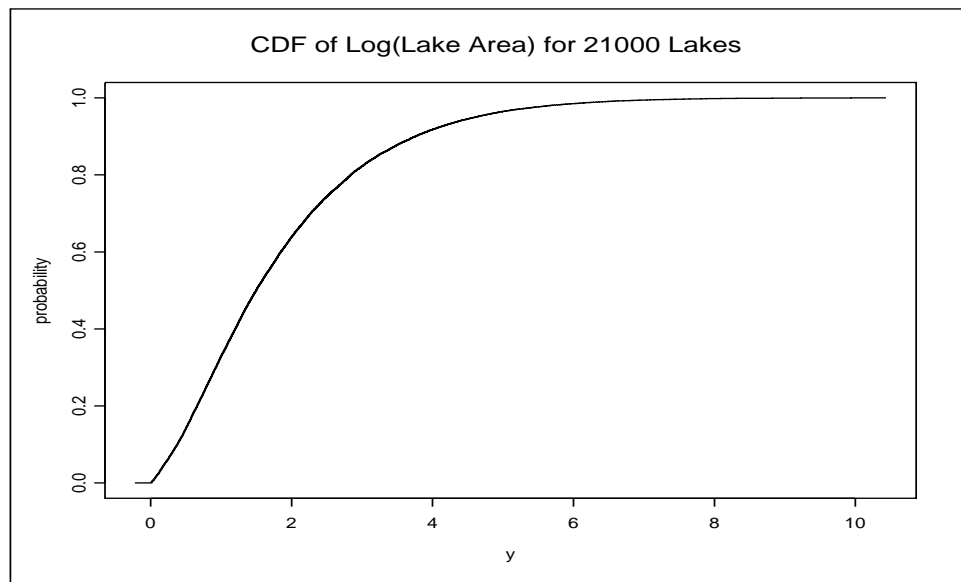
$$F(t) = \frac{1}{N} \sum_{i \in U} I_{\{Y_i \leq t\}}$$

- Some Notation:

finite population: $U = \{1, 2, \dots, N\}$

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

	Parametric	Nonparametric
model based	Chambers and Dunstan	Dorfman
model assisted	Rao, Kovar, Mantel	LPR

- Model:

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

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

$$\hat{F}_{RKM}(t) = \underbrace{\frac{1}{N} \sum_{i \in U} \hat{G}_i}_{\text{model-based prediction}} + \underbrace{\sum_{i \in s} \frac{I_{\{Y_j \leq t\}} - \hat{G}_{ic}}{N\pi_i}}_{\text{design-bias adjustment}}$$

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 $I_{\{Y_i \leq t\}}$:

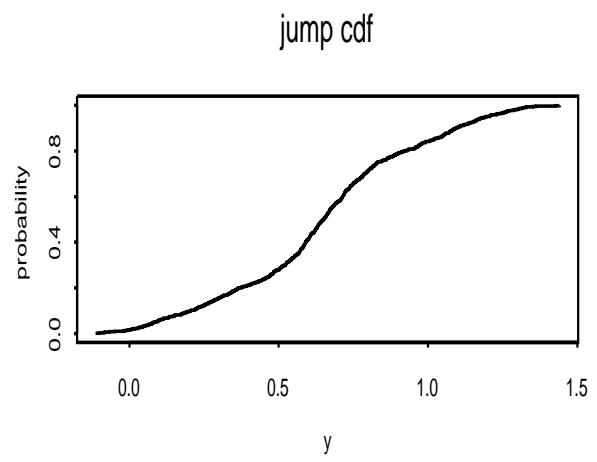
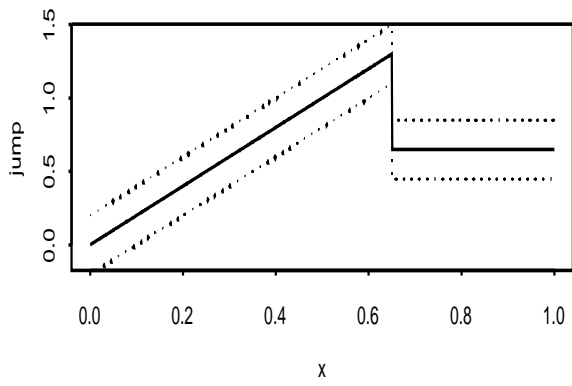
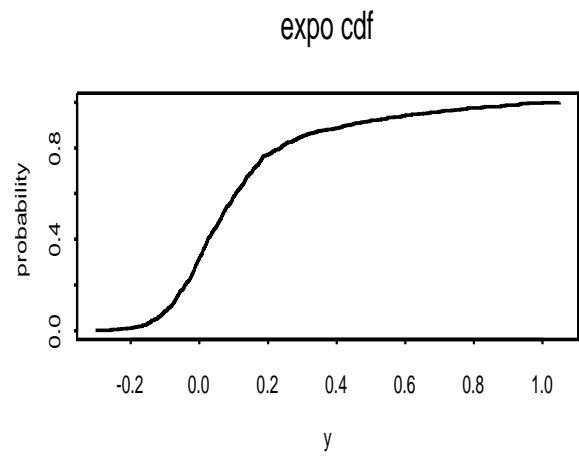
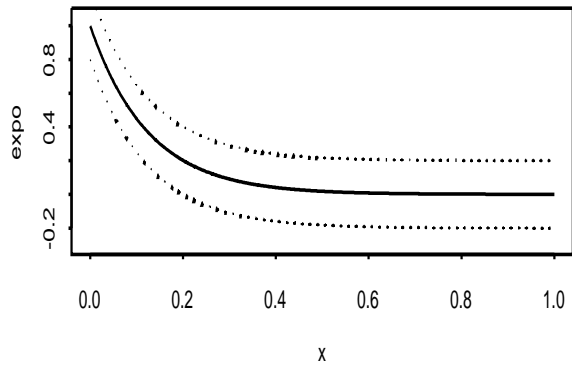
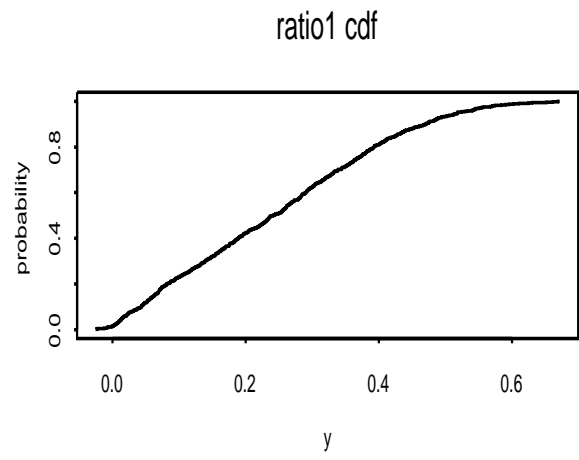
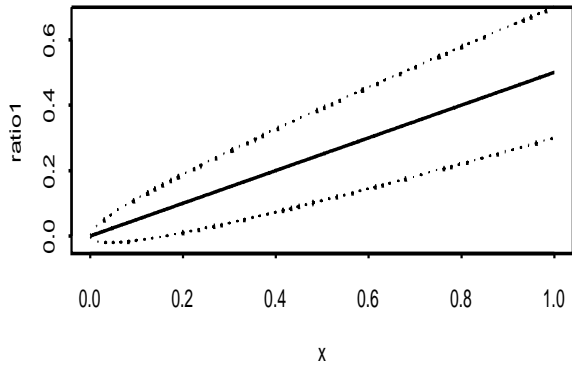
$$\hat{F}_{LPR}(t) = \sum_{i \in U} \hat{\mu}_i + \sum_{i \in s} \frac{I_{\{Y_i \leq t\}} - \hat{\mu}_i}{\pi_i}$$

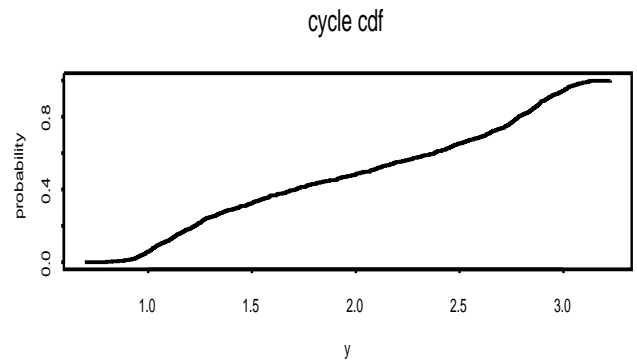
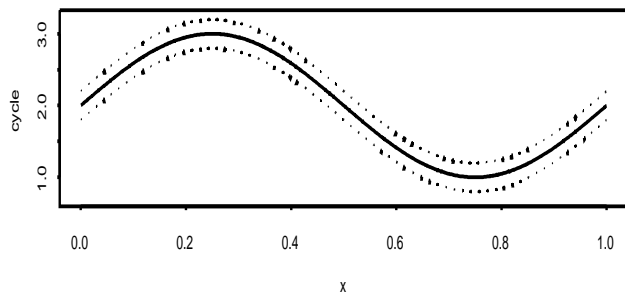
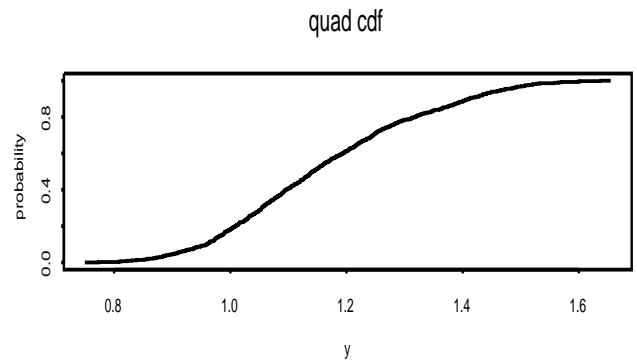
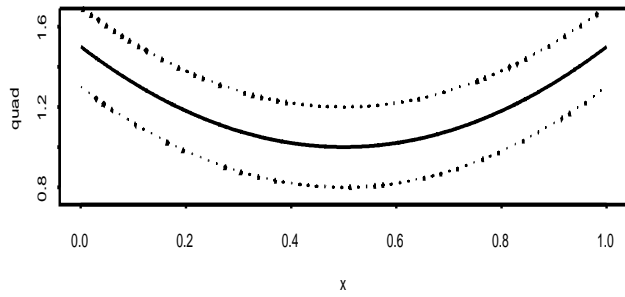
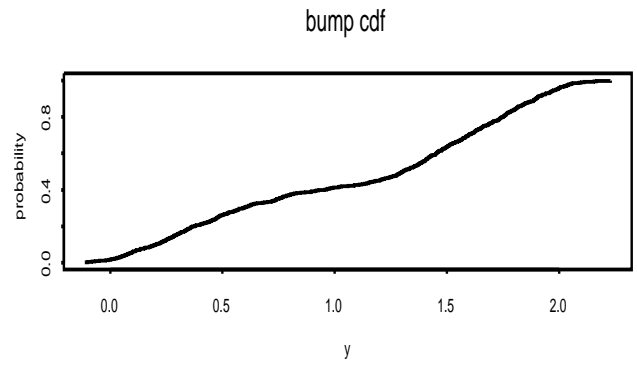
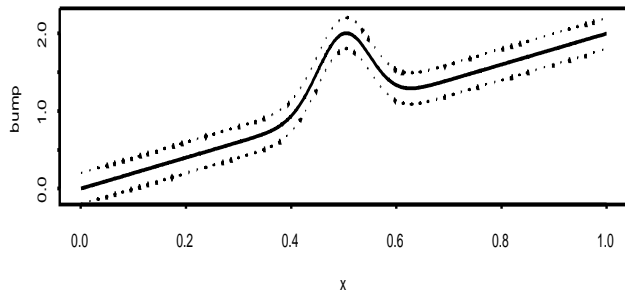
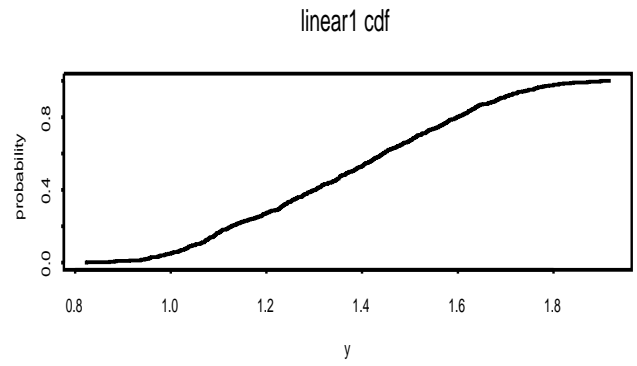
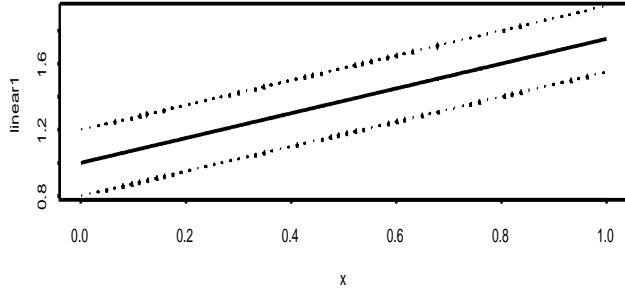
where μ is a new smooth function of x_i :

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

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

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

NOTE:

m(x) not misspecified

m(x) misspecified

Summary and Further Work

	Parametric	Nonparametric
model based	Chambers and Dunstan	Dorfman
model assisted	Rao, Kovar, Mantel	LPR

- variance misspecification bias
- quantile estimation
- analytical comparisons
(Chambers, Dorfman, Hall (1992))

Percent Relative Bias Results

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

References

- Breidt, F. J., Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Ann. Statist.* **28**, 1026-1053.
- Chambers, R. L., Dorfman, A. H., Hall, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika* **79**, 577-82.
- Chambers, R. L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika* **73**, 597-604.
- Dorfman, A. H. (1992). Nonparametric regression for estimating totals in finite populations. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 622-625.
- Rao, J. N. K., Kovar, J. G., Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* **77**, 365-75.