

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 $U = \{1, 2, \dots, N\}$ (“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

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

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

$$\hat{F}_{CD}(t) = \underbrace{\frac{1}{N} \sum_{i \in U-s} \hat{G}_i}_{\substack{\text{model-based} \\ \text{prediction}}} + \underbrace{\frac{1}{N} \sum_{i \in s} I_{\{y_i \leq t\}}}_{\substack{\text{sampled} \\ \text{values}}}$$

- \hat{G}_i estimates $G\left(\frac{t-m(x_i)}{v^{1/2}(x_i)}\right) = E_m(I_{\{y_i \leq t\}})$

- asymptotically unbiased when $m(x_i)$ and $v(x_i)$ correctly specified

Model-Assisted Parametric: RKM Estimator

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

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

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

- asymptotically design and model unbiased

Motivation for Nonparametric Methods

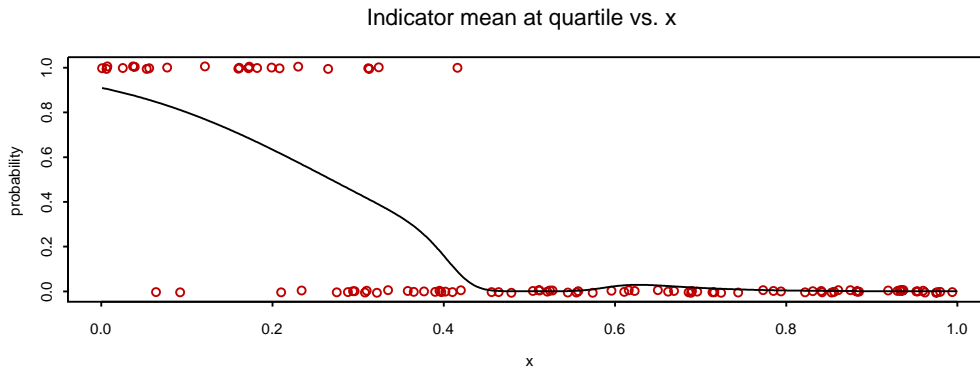
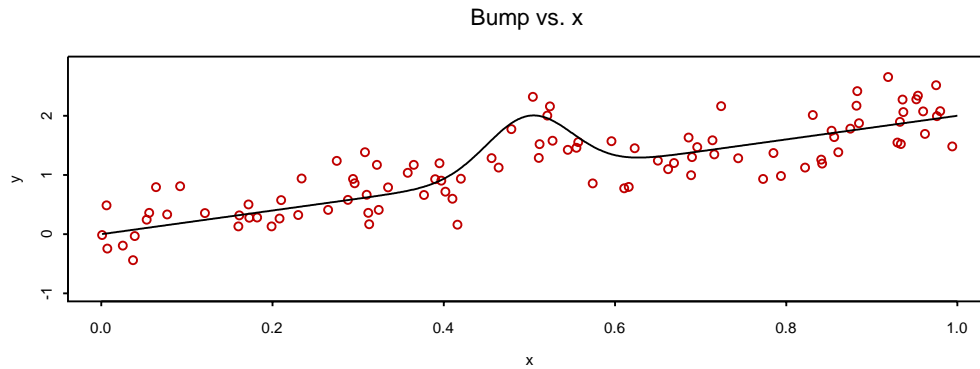
- $E_m I_{\{y_i \leq t\}} = \Pr \{y_i \leq t\} = G(v^{-1/2}(x_i)(t - m(x_i)))$
- Mean function misspecification
 - CD will be biased
 - RKM will be inefficient
 - nonparametric methods only assume $m(x_i)$ is smooth
- Variance misspecification
 - CD and RKM assume $v(x_i)$ is known
 - nonparametric assumes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

$$\mathbf{X}_{si} = \left[1 \quad x_j - x_i \quad \cdots \quad (x_j - x_i)^q \right]_{j \in s}$$

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

$$\mathbf{W}_{si} = \text{diag} \left\{ \frac{1}{\pi_j h} K \left(\frac{x_j - x_i}{h} \right) \right\}_{j \in s}$$

- Sample smoother vector at x_i :

$$\mathbf{s}'_{si} = [1 \ 0 \ \cdots \ 0] (\mathbf{X}'_{si} \mathbf{W}_{si} \mathbf{X}_{si})^{-1} \mathbf{X}'_{si} \mathbf{W}_{si}$$

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

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

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 $\mathbf{I}_s = [I_{\{y_i \leq t\}}]_{i \in s}$
- Estimate $\hat{G}(v^{-1/2}(x_i)(t - m(x_i)))$ by $\hat{g}_i = \mathbf{s}'_{si} \mathbf{I}_s$
- Then

$$\hat{F}_{LPR}(t) = \frac{1}{N} \sum_{i \in U} \hat{g}_i + \frac{1}{N} \sum_{i \in s} \frac{I_{\{y_i \leq t\}} - \hat{g}_i}{\pi_i}$$

- Can construct model-based nonparametric estimator analogously

Properties of LPR CDF Estimator

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

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

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:

estimate of sum = sum of the estimates

$$\sum_{i \in s} \omega_{is}(y_i + z_i) = \sum_{i \in s} \omega_{is}y_i + \sum_{i \in s} \omega_{is}z_i$$

- Mean of LPR-estimated cdf is LPR-estimated mean

$$\int y d\hat{F}_{LPR}(y) = \sum_{i \in s} \omega_{is}y_i = \frac{\hat{T}_{LPR}}{N}$$

– assuming weights are non-negative

Range of Estimators

- From specific to generic:

Specific	Model-based parametric	CD
↓	Model-assisted parametric	RKM
↓	Model-based nonparametric	DORF
↓	Model-assisted nonparametric	LPR
Generic	Design-based	HT

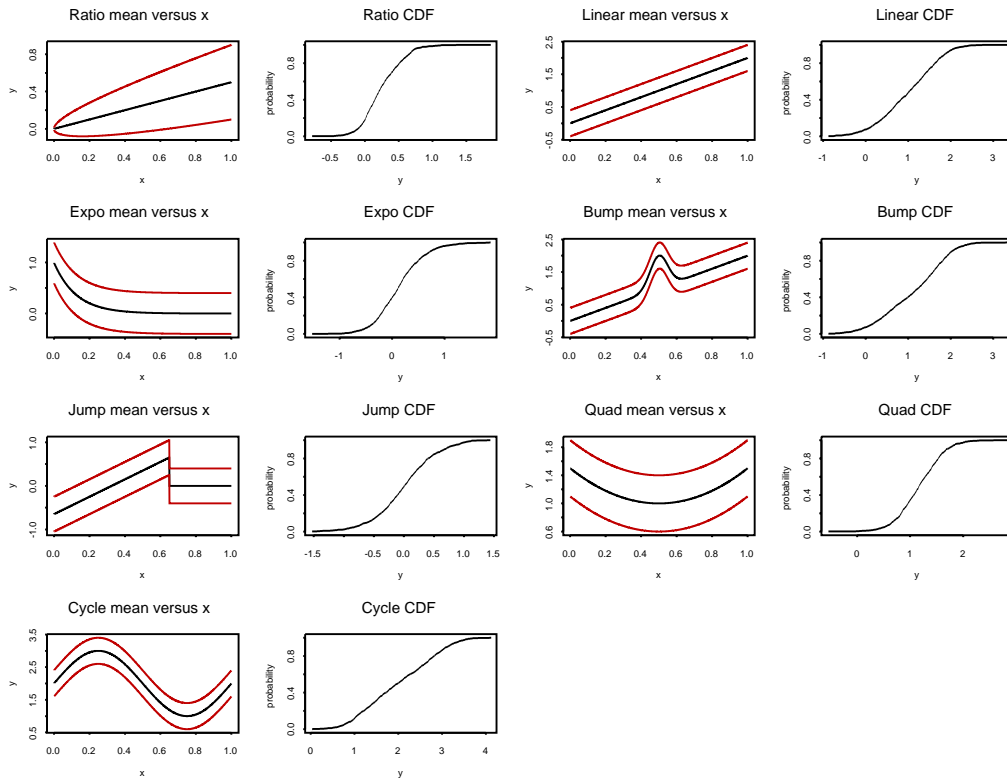
CDF Simulation Study Design

- Compare seven estimators via simulation:

Type	Estimator	Model
design-based	HT	no model
model-based	CD0	$\beta_1 x + x^{1/2} \epsilon$
	CD1	$\beta_0 + \beta_1 x + \epsilon$
	DORF	$m(x) + v^{1/2}(x) \epsilon$
model-assisted	RKM0	$\beta_1 x + x^{1/2} \epsilon$
	RKM1	$\beta_0 + \beta_1 x + \epsilon$
	LPR	$m(x) + v^{1/2}(x) \epsilon$

Simulated Response Variables

- 7 response variables with $x_i \sim \text{Unif}(0, 1)$, $\epsilon_i \sim N(0, \sigma^2)$



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)

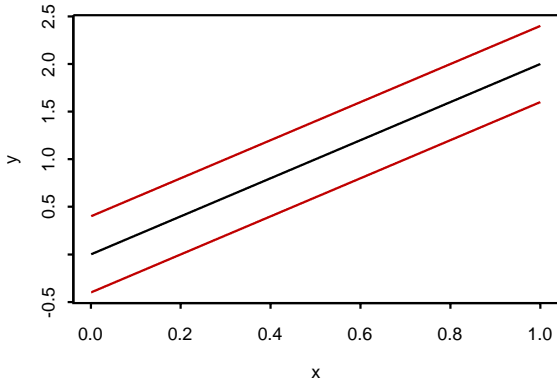
$$\frac{MSE(\hat{F}_*(t))}{MSE(\hat{F}_{LPR}(t))}$$

- Return percent relative biases:

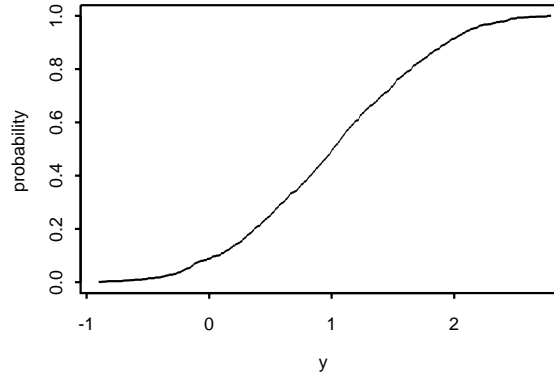
$$\left(\overline{(\hat{F}(t) - F(t))} / F(t) \right) 100\%$$

Smoothing to Estimate the Linear CDF

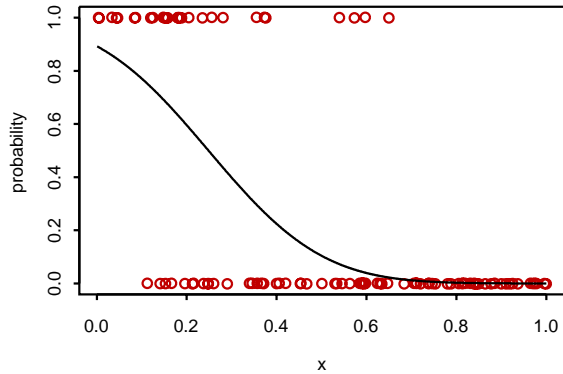
Linear mean versus x



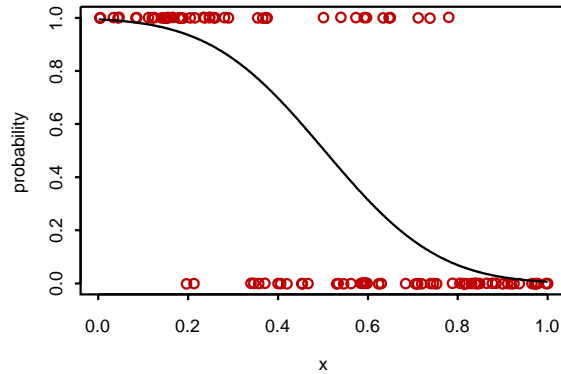
Linear CDF



Indicator mean at quartile vs. x

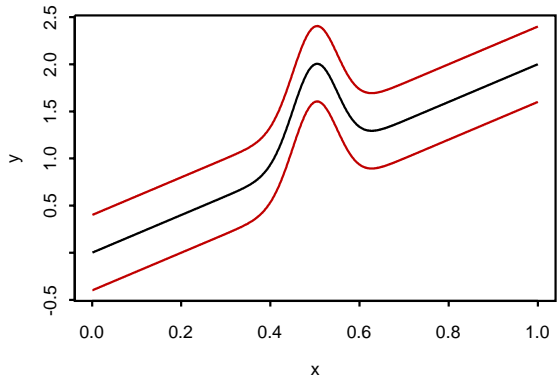


Indicator mean at median vs. x

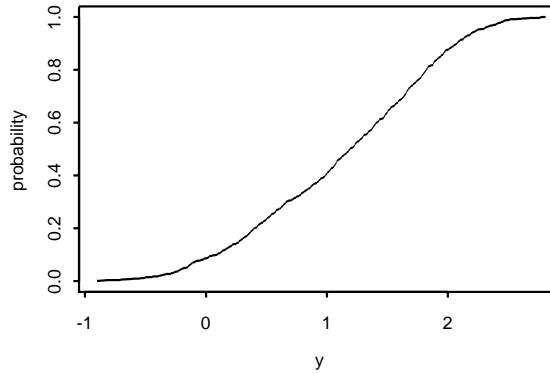


Smoothing to Estimate the Bump CDF

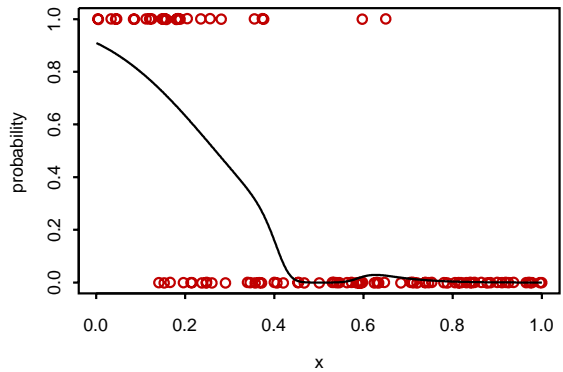
Bump mean versus x



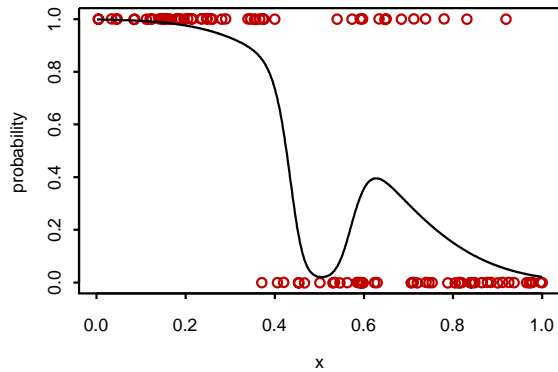
Bump CDF



Indicator mean at quartile vs. x

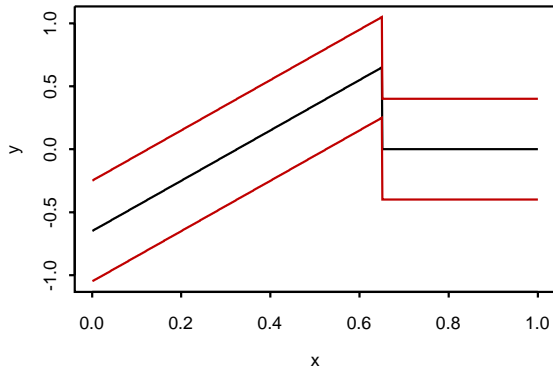


Indicator mean at median vs. x

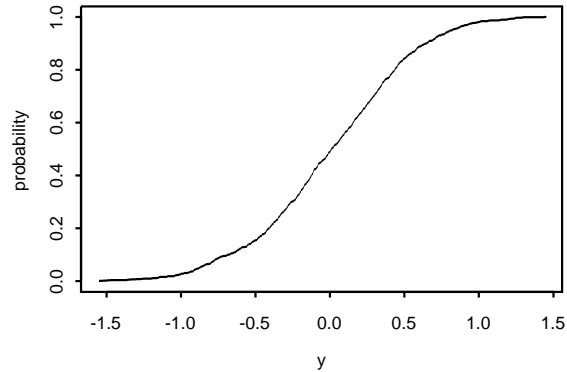


Smoothing to Estimate the Jump CDF

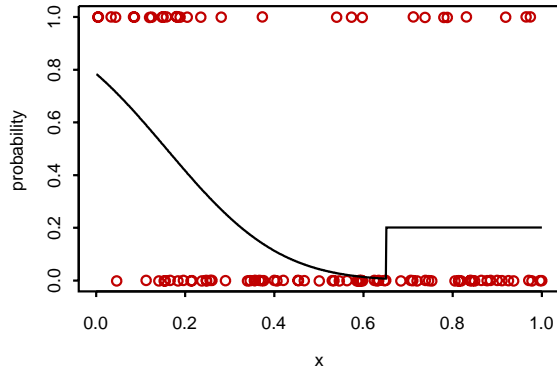
Jump mean versus x



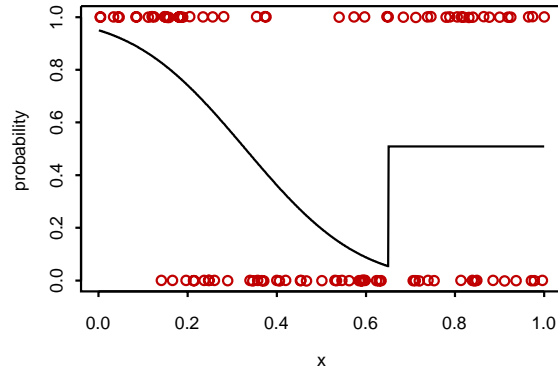
Jump CDF



Indicator mean at quartile vs. x



Indicator mean at median vs. x



CDF Simulation Results: Bias

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

Design-based	HT	< 0.5% RB
Model-assisted	RKM0, RKM1, LPR	< 2% RB
Model-based par.	CD0, CD1	usually > 2% RB
Model-based nonpar.	DORF	several 2%–6% RB

- Design bias adjustment works!

CDF Simulation Numerical Results

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

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

– $m(x)$ not misspecified

– $m(x)$ misspecified

MSE Comparisons to Generic Estimators

- LPR dominates HT and DORF in nearly all cases

	Min	Q1	Med	Q3	Max
HT	0.91	1.22	2.01	3.16	10.25
DORF	0.98	1.13	1.38	1.63	3.29

- LPR is competitive with RKM0, RKM1

	Min	Q1	Med	Q3	Max
RKM0	0.69	0.94	1.16	1.85	16.55
RKM1	0.69	0.96	1.03	1.37	4.65

MSE Comparisons for Specific Estimators

- **Mean correct, variance correct**

	$\sigma = 0.1$	$\sigma = 0.4$
CD0	0.14, 0.14, 0.39, 0.39	0.56, 0.60, 0.67, 0.71
CD1	0.17, 0.19, 0.19, 0.23	0.54, 0.56, 0.58, 0.60

- **Mean correct, variance incorrect**

	$\sigma = 0.1$	$\sigma = 0.4$
CD0	0.07, 0.08, 0.55, 0.67	1.15, 1.21, 2.74, 2.86
CD1	0.27, 0.27, 0.74, 0.75	0.71, 0.77, 1.85, 1.94

- **Mean incorrect**

	Min	Q1	Med	Q3	Max
CD0	0.48	1.01	2.71	14.98	84.39
CD1	0.64	0.91	1.44	3.68	18.36

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, $h = 0.25$, $\sigma = 0.4$

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

- 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