

Estimating Distribution Functions from Survey Data Using Nonparametric Regression

Alicia Johnson

Department of Statistics
Colorado State University

The work reported here was developed under the STAR Research Assistance Agreements CR-829095 and R-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 authors and STARMAP. EPA does not endorse any products or commercial services mentioned in this report.

Outline of Research

- Introduction to CDF estimation
 - use of auxiliary information
 - parametric methods
 - nonparametric methods
- Nonparametric CDF estimation
 - local polynomial regression
 - model-assisted estimator
- Simulation results
 - Monte Carlo comparison of several CDF estimators
 - Monte Carlo comparison of several quantile estimators
- Semiparametric CDF estimation
 - use of multiple auxiliary variables
 - semiparametric regression
 - model-assisted estimator
- Empirical example

Introduction Outline

- Finite population CDF estimation for y
- Horvitz-Thompson estimator
- Estimation with auxiliary information
 - auxiliary information \mathbf{x} available for entire landscape
 - parametric models relating y to \mathbf{x}
 - * Chambers and Dunstan estimator (1986)
 - * Rao, Kovar, Mantel estimator (1990)
- Motivation for nonparametric methods

Finite Population CDF Estimation

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

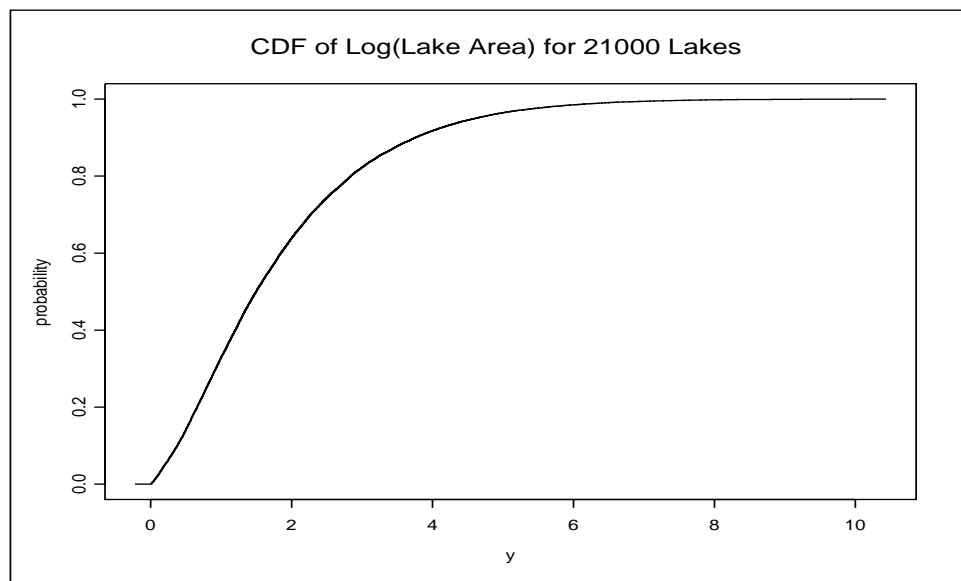
- Some Notation:

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

y_i observed for sample: $A \subset U$ of size n

$\pi_i = \Pr \{i \in A\}$

$\pi_{ij} = \Pr \{i, j \in A\}$



Horvitz-Thompson Estimator

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

- Design unbiased
- No dependence on any model
- Does not incorporate auxiliary information \mathbf{x}
- How do we incorporate \mathbf{x} for the entire landscape?

Estimation with Auxiliary Information

- x_i known for all $i \in U$
- Superpopulation model:

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

where:

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

- Superpopulation model allows inferences to be made about nonsampled portion of y

Parametric Methods

- Parametric superpopulation:

$$y_i = \beta_0 + \beta_1 x_i + v^{1/2}(x_i)\epsilon_i$$

- Assumptions:
 - $v^{1/2}(x_i)$ is known and strictly positive
 - mean function, $m(x_i)$, is linear

Parametric Methods Continued

- CD estimator

- Chambers and Dunstan (1986)

- model-based

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

- \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 $v^{1/2}(x_i)$ correctly specified

Parametric Methods Continued

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

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

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

- asymptotically design and model unbiased

Motivation for Nonparametric Methods

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

- mean function misspecification bias
 - CD and RKM assume $m(x_i) = \beta_0 + \beta_1 x_i$
 - 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
 - nonparametric methods only assume $v^{1/2}(x_i)$ is smooth and strictly positive

Nonparametric CDF Estimation Outline

- Nonparametric regression
 - general overview
 - bandwidth selection
- Local polynomial regression (LPR)
 - overview
 - application to survey sampling estimation
 - application to finite population CDF estimation

Nonparametric Regression Overview

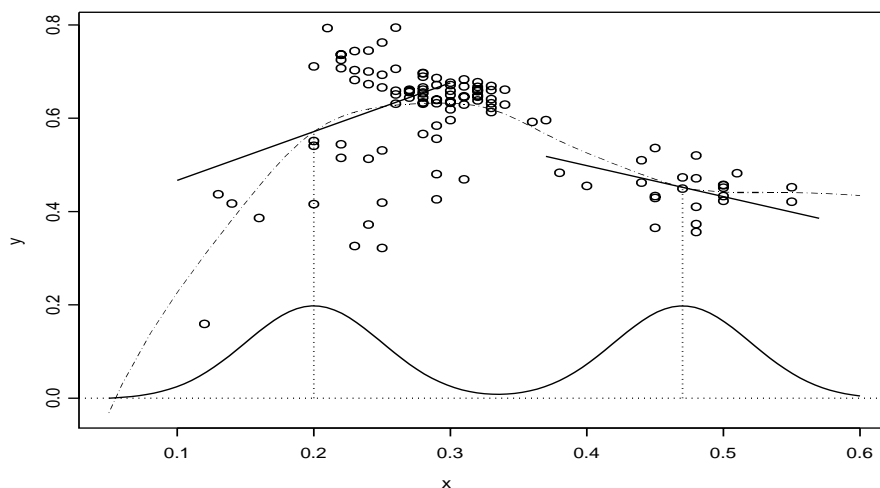
Recall superpopulation model:

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

- Assume $v^{1/2}(x_i)$ is strictly positive
- $m(x)$ is locally approximated
 - place more weight on y_i 's corresponding to x_i 's close to x
 - weights are determined from $\frac{1}{h}K\left(\frac{x_i-x}{h}\right)$ where:
 - * $K(\bullet)$ is a kernel function centered at x
 - * $K(\bullet)$ is scaled by bandwidth h

Nonparametric Regression Overview Continued

- Bandwidth selection
 - large bandwidth
 - * increases spread of $K(\bullet)$
 - * may result in oversmoothing
 - small bandwidth
 - * decreases spread of $K(\bullet)$
 - * may result in undersmoothing



Local Polynomial Regression

- $m(x)$ is locally approximated by a polynomial of degree q

- Sample design matrix ($n \times (q + 1)$):

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

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

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

- Sample smoother vector at x_i :

$$\mathbf{s}'_{Ai} = \mathbf{e}'_1 (\mathbf{X}'_{Ai} \mathbf{W}_{Ai} \mathbf{X}_{Ai})^{-1} \mathbf{X}'_{Ai} \mathbf{W}_{Ai}$$

where $\mathbf{e}'_1 = [1 \ 0 \ \cdots \ 0]_{1 \times (q+1)}$

Local Polynomial Regression Continued

- Define $\mathbf{y}_A = [y_i]_{i \in A}$
- $\hat{m}_i = \mathbf{s}'_{Ai} \mathbf{y}_A$

\hat{m}_i is the sample estimate of the population kernel smooth LPR fit at x_i

LPR in Survey Sampling Estimation

Finite population total: $T_y = \sum_{i \in U} y_i$

- Breidt and Opsomer (2000) LPR estimator for population total:

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

where $\hat{m}_i = \mathbf{s}'_{Ai} \mathbf{y}_A$

- Design properties of \hat{T}_{LPR} :
 - nonparametric, model-assisted
 - design consistent
 - asymptotically design unbiased
- Estimated variance of \hat{T}_{LPR} :

$$\widehat{\text{Var}}(\hat{T}_{LPR}) = \sum_{i,j \in A} (y_i - \hat{m}_i)(y_j - \hat{m}_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{1}{\pi_{ij}}$$

Local Polynomial Regression CDF Estimator

- Based on \hat{T}_{LPR}

- Define $\mathbf{I}_A = [I_{\{y_i \leq t\}}]_{i \in A}$

- Replace y_i in \hat{T}_{LPR} with $I_{\{y_i \leq t\}}$:

$$\hat{F}_{LPR}(t) = \frac{1}{N} \left[\sum_{i \in U} \hat{m}_i + \sum_{i \in A} \frac{I_{\{y_i \leq t\}} - \hat{m}_i}{\pi_i} \right]$$

where $\hat{m}_i = \mathbf{s}'_{Ai} \mathbf{I}_A$

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

Local Polynomial Regression CDF Estimator Continued

- Design properties of \hat{T}_{LPR} hold for \hat{F}_{LPR} :
 - nonparametric, model-assisted
 - design consistent
 - asymptotically design unbiased

- Estimated variance of $\hat{F}_{LPR}(t)$:

$$\widehat{\text{Var}}(\hat{F}_{LPR}(t)) = \frac{1}{N^2} \sum_{i,j \in A} (I_{ti} - \hat{m}_i)(I_{tj} - \hat{m}_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{1}{\pi_{ij}}$$

where $I_{tk} = I_{\{y_k \leq t\}}$

Simulation Results Outline

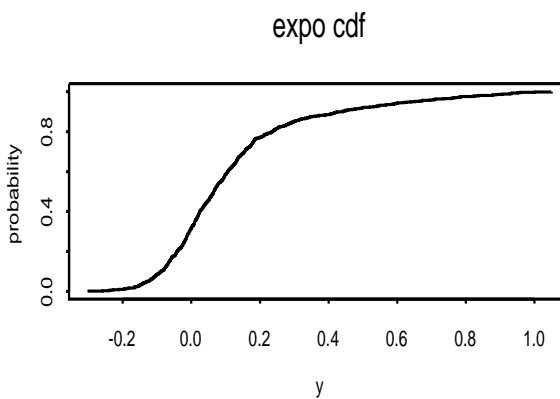
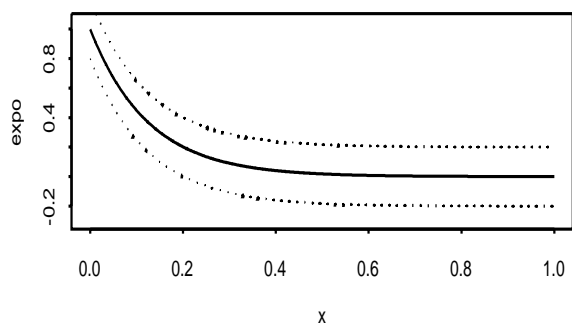
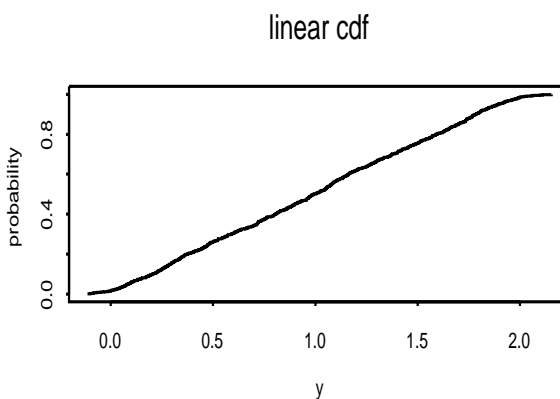
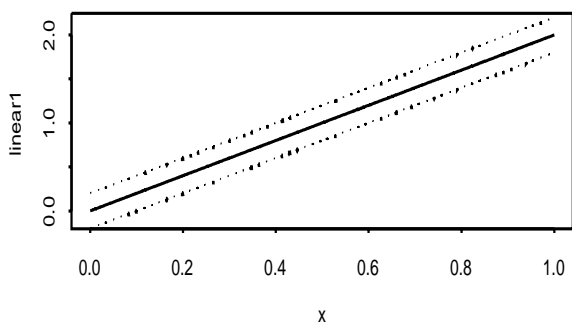
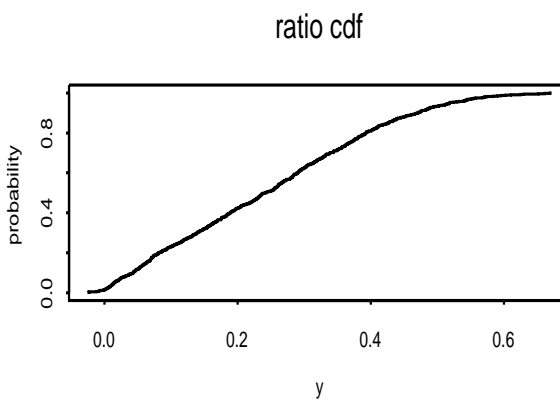
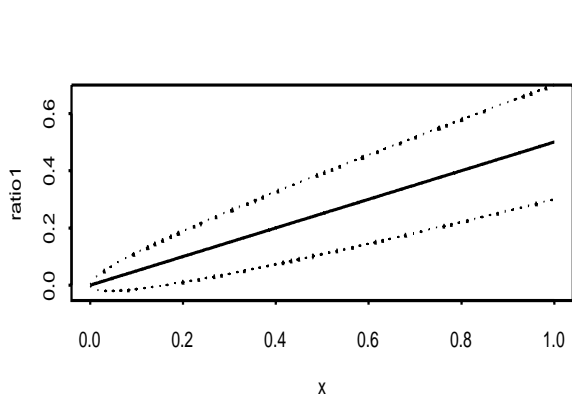
- CDF simulation
 - study design
 - numerical results
- Quantile simulation
 - quantile estimation
 - study design
 - numerical results

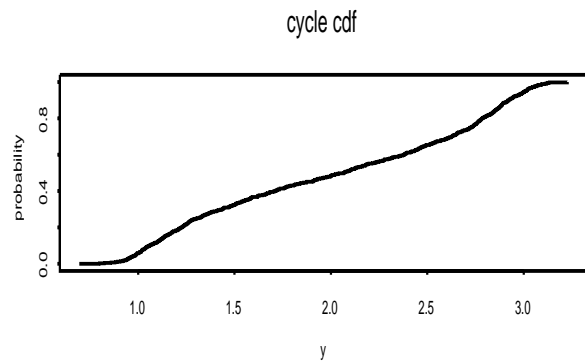
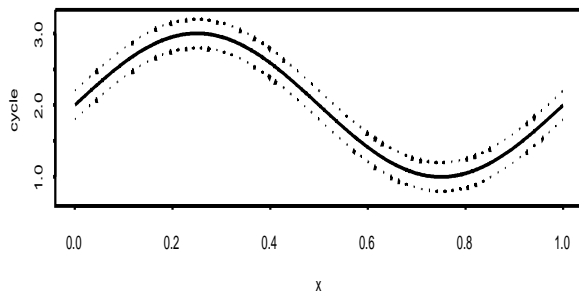
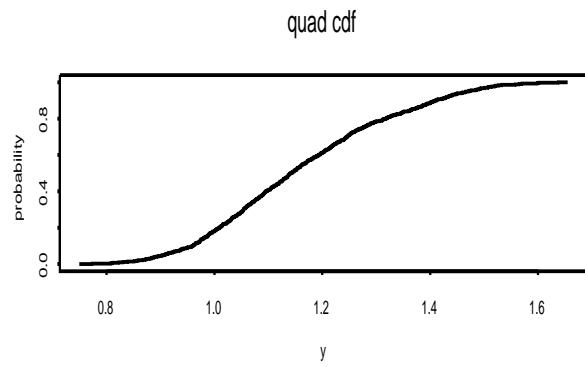
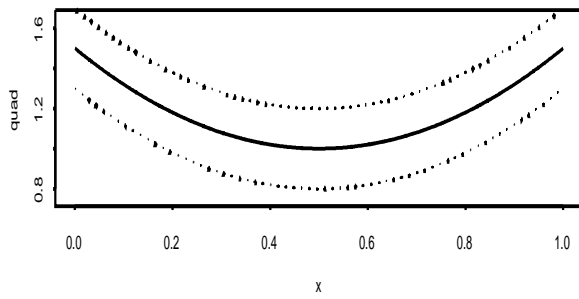
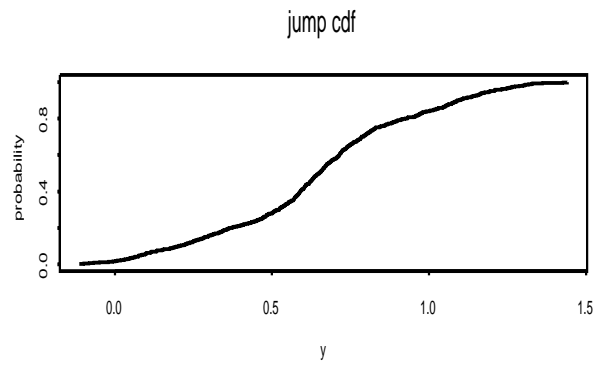
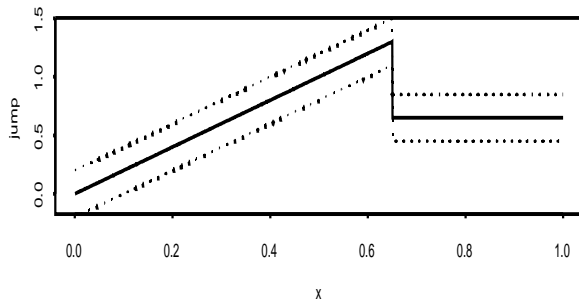
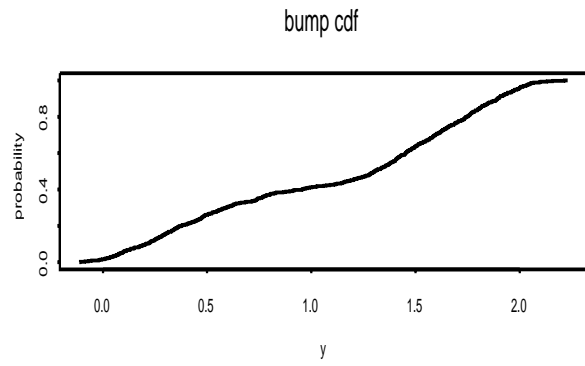
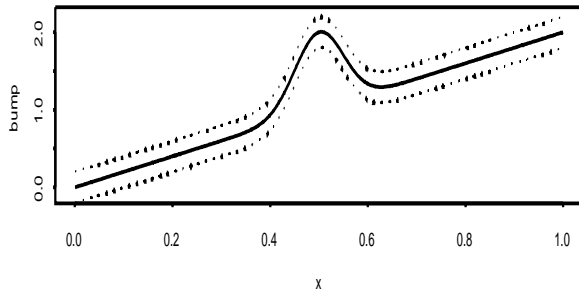
	Parametric	Nonparametric
model-based	Chambers and Dunstan	Dorfman
model-assisted	Rao, Kovar, Mantel	LPR

CDF Simulation Study Design

- Estimators:
 - HT
 - CD0 ($\beta_0 = 0, v^{1/2}(x) = x^{1/2}$)
 - CD1 ($v^{1/2}(x) = 1$)
 - RKM0 ($\beta_0 = 0, v^{1/2}(x) = x^{1/2}$)
 - RKM1 ($v^{1/2}(x) = 1$)
 - LPR ($q = 1$)
 - DORF
- 7 populations generated from $x_i \sim \text{Unif}(0, 1)$
and $\epsilon_i \sim \text{N}(0, \sigma^2)$
- Model misspecification
- Simple random sampling ($\pi_i = \frac{n}{N}$)

Graphs of simulated populations and their cdfs:





CDF Simulation Study Design Continued

- $N = 1000, n = 100$
- 1000 reps
- \hat{F}_{LPR} calculated using Epanechnikov kernel function:

$$K(x) = \frac{3}{4}(1 - x^2)$$

- Return MSE ratios: (> 1 favors LPR)

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

- Return percent relative biases:

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

where α is the true value of $F(t)$

Table 1: MSE ratios for CDF estimation at the median

Population	h	σ	HT	CD0	CD1	RKM0	RKM1	DORF
Ratio	0.10	0.1	2.72	0.39	0.27	0.92	0.94	1.51
	0.10	0.4	1.18	0.67	1.85	0.91	0.92	1.15
	0.25	0.1	2.77	0.39	0.27	0.94	0.96	1.56
	0.25	0.4	1.24	0.71	1.94	0.95	0.97	1.22
Linear	0.10	0.1	7.02	0.67	0.23	0.85	0.84	2.48
	0.10	0.4	2.07	2.74	0.54	0.93	0.93	1.32
	0.25	0.1	5.73	0.55	0.19	0.69	0.69	2.32
	0.25	0.4	2.16	2.86	0.56	0.97	0.97	1.40
Expo	0.10	0.1	1.35	2.00	3.59	1.80	1.01	1.11
	0.10	0.4	1.00	0.96	0.79	1.14	0.94	1.08
	0.25	0.1	1.42	2.10	3.77	1.90	1.06	1.23
	0.25	0.4	1.06	1.02	0.83	1.20	0.99	1.17
Bump	0.10	0.1	4.85	19.47	17.51	2.15	2.37	1.96
	0.10	0.4	2.58	7.24	2.98	1.23	1.29	1.41
	0.25	0.1	3.05	12.26	11.02	1.35	1.49	1.49
	0.25	0.4	2.26	6.36	2.62	1.08	1.14	1.39
Jump	0.10	0.1	1.99	0.77	2.38	1.51	1.90	1.39
	0.10	0.4	1.20	1.19	0.90	1.07	1.11	1.13
	0.25	0.1	1.87	0.72	2.24	1.42	1.79	1.49
	0.25	0.4	1.26	1.26	0.95	1.13	1.18	1.24
Quad	0.10	0.1	2.09	0.70	2.00	2.67	2.11	1.41
	0.10	0.4	0.97	0.48	0.91	1.26	0.98	1.09
	0.25	0.1	2.03	0.68	1.95	2.60	2.06	1.89
	0.25	0.4	1.04	0.51	0.97	1.34	1.05	1.20
Cycle	0.10	0.1	10.25	13.67	4.66	16.55	3.37	3.29
	0.10	0.4	2.79	3.11	1.18	4.29	1.37	1.47
	0.25	0.1	5.79	7.73	2.63	9.35	1.90	2.42
	0.25	0.4	2.79	3.11	1.19	4.29	1.38	1.57

Table 2: Percent relative biases for cdf estimators at the median

Population	h	σ	HT	CD0	CD1	RKM0	RKM1	LPR	DORF
Ratio	0.10	0.1	0.29	2.13	-1.45	0.13	0.11	-0.96	-0.90
	0.10	0.4	0.42	2.12	-9.53	0.33	0.22	-0.55	-0.64
	0.25	0.1	0.29	2.13	-1.45	0.13	0.11	-0.89	-0.64
	0.25	0.4	0.42	2.12	-9.53	0.33	0.22	-0.48	-1.01
Linear	0.10	0.1	0.16	2.61	1.48	0.04	0.03	-1.04	-0.96
	0.10	0.4	0.35	10.01	3.27	0.26	0.22	-0.85	-0.80
	0.25	0.1	0.16	2.61	1.48	0.04	0.03	-1.02	-0.85
	0.25	0.4	0.35	10.01	3.27	0.26	0.22	-0.83	-0.62
Expo	0.10	0.1	-0.25	8.01	-14.56	-0.35	-0.22	0.49	0.56
	0.10	0.4	-0.05	4.05	-2.98	-0.15	0.07	0.40	0.72
	0.25	0.1	-0.25	8.01	-14.56	-0.35	-0.22	0.56	1.07
	0.25	0.4	-0.05	4.05	-2.98	-0.15	0.07	0.51	1.54
Bump	0.10	0.1	0.45	18.86	17.84	0.53	0.52	-0.79	-0.59
	0.10	0.4	0.28	15.02	9.11	0.24	0.17	-1.04	-0.95
	0.25	0.1	0.45	18.86	17.84	0.53	0.52	-0.80	-0.55
	0.25	0.4	0.28	15.02	9.11	0.24	0.17	-0.97	-0.49
Jump	0.10	0.1	0.18	2.80	6.88	0.01	-0.18	-0.66	-0.84
	0.10	0.4	0.29	6.87	1.56	0.19	-0.04	-0.42	-0.60
	0.25	0.1	0.18	2.80	6.88	0.01	-0.18	-0.45	-0.84
	0.25	0.4	0.29	6.87	1.56	0.19	-0.04	-0.22	-1.06
Quad	0.10	0.1	0.27	0.21	-0.10	0.12	0.74	0.36	0.35
	0.10	0.4	0.40	2.49	-0.35	0.29	0.58	0.32	0.86
	0.25	0.1	0.27	0.21	-0.10	0.12	0.74	0.08	2.91
	0.25	0.4	0.40	2.49	-0.35	0.29	0.58	0.41	1.90
Cycle	0.10	0.1	-0.01	7.34	4.49	-0.07	0.21	0.65	0.53
	0.10	0.4	-0.23	6.15	2.95	-0.30	-0.07	-0.01	0.24
	0.25	0.1	-0.01	7.34	4.49	-0.07	0.21	0.91	0.44
	0.25	0.4	-0.23	6.15	2.95	-0.30	-0.07	0.24	0.15

CDF Simulation Numerical Results

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

Population	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

NOTE:

m(x) not misspecified

m(x) misspecified

CDF Simulation Results

- LPR more efficient than HT and DORF

	Minimum MSE ratio	Maximum MSE ratio
HT	1.04	2.79
DORF	1.17	1.57

- CD and RKM misspecify mean function:
 - LPR is competitive or more efficient

	Minimum MSE ratio	Maximum MSE ratio
CD0	0.51	6.36
CD1	0.83	2.62
RKM0	0.97	4.29
RKM1	0.99	1.38

- CD0 MSE ratio of 0.51 is for “quad”
- next smallest CD0 MSE ratio is 1.02

CDF Simulation Results Continued

- Linear mean function is appropriate

- LPR competitive with RKM

$$0.95 \leq \frac{MSE(\hat{F}_{RKM*}(t))}{MSE(\hat{F}_{LPR}(t))} \leq 0.97$$

- LPR is less efficient than CD when CD correctly specifies both $m(x)$ and $v^{1/2}(x)$

- * CD0 MSE ratio for “ratio” is 0.71

- * CD1 MSE ratio for “linear” is 0.56

- LPR more efficient than CD1 when CD1 misspecifies $v^{1/2}(x)$ for “ratio” population (MSE ratio is 1.94)

Effect of Bandwidth

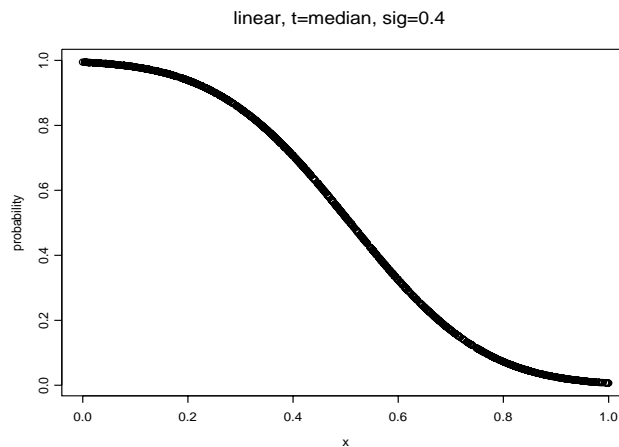
- How does selection of h affect LPR efficiency?
- Complicated problem
- Need to look at what we are smoothing:

$$\begin{aligned} G\left(\frac{t-m(x_i)}{v^{1/2}(x_i)}\right) &= E_m I_{y_i \leq t} \\ &= P(y_i \leq t) \\ &= P\left(\epsilon_i \leq \frac{t-m(x_i)}{v^{1/2}(x_i)}\right) \\ &= \Phi\left(\frac{t-m(x_i)}{v^{1/2}(x_i)}\right) \end{aligned}$$

- Shape of $\Phi(\bullet)$ depends on population and selection of σ and t

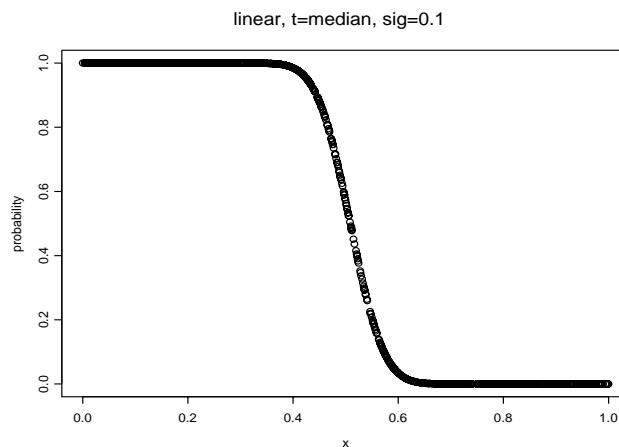
Effect of Bandwidth: Linear Population

- $\Phi\left(\frac{t-m(x_i)}{v^{1/2}(x_i)}\right) = \Phi(\text{median} - (1 + 2(x - 0.5))), \epsilon_i \sim N(0, 0.4^2)$



h	σ	HT	CD0	CD1	RKM0	RKM1	DORF
0.10	0.4	2.07	2.74	0.54	0.93	0.93	1.32
0.25	0.4	2.16	2.86	0.56	0.97	0.97	1.40

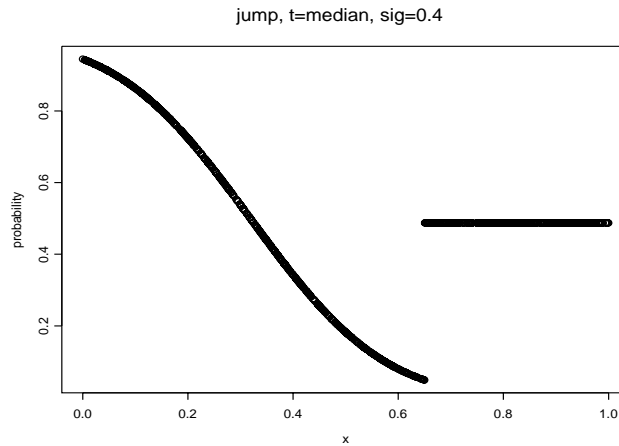
- $\Phi\left(\frac{t-m(x_i)}{v^{1/2}(x_i)}\right) = \Phi(\text{median} - (1 + 2(x - 0.5))), \epsilon_i \sim N(0, 0.1^2)$



h	σ	HT	CD0	CD1	RKM0	RKM1	DORF
0.10	0.1	7.02	0.67	0.23	0.85	0.84	2.48
0.25	0.1	5.73	0.55	0.19	0.69	0.69	2.32

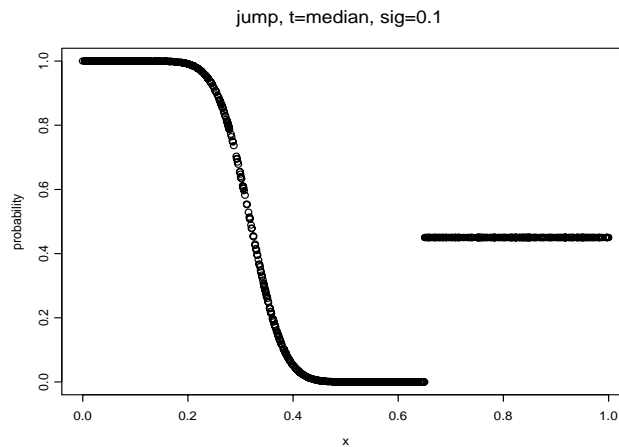
Effect of Bandwidth: Jump Population

- $\Phi\left(\frac{t-m(x_i)}{v^{1/2}(x_i)}\right) = \Phi(\text{median} - (0.35 + 2(x - 0.5))I_{x \leq 0.65}), \epsilon_i \sim N(0, 0.4^2)$



h	σ	HT	CD0	CD1	RKM0	RKM1	DORF
0.10	0.4	1.20	1.19	0.90	1.07	1.11	1.13
0.25	0.4	1.26	1.26	0.95	1.13	1.18	1.24

- $\Phi\left(\frac{t-m(x_i)}{v^{1/2}(x_i)}\right) = \Phi(\text{median} - (0.35 + 2(x - 0.5))I_{x \leq 0.65}), \epsilon_i \sim N(0, 0.1^2)$



h	σ	HT	CD0	CD1	RKM0	RKM1	DORF
0.10	0.1	1.99	0.77	2.38	1.51	1.90	1.39
0.25	0.1	1.87	0.72	2.24	1.42	1.79	1.49

CDF Simulation Results Continued

- Percent relative bias
 - up to 15 percent for model-based estimators (CD & DORF)
 - less than 1 percent for model-assisted estimators (LPR & RKM)

$$0.15 \leq |\% \text{rel. bias}| \leq 15.02$$

$$0.04 \leq |\% \text{rel. bias}| \leq 0.97$$

Quantile Estimation Simulation

$$\theta(\alpha) = \min\{t : F(t) \geq \alpha\}$$

- $\hat{\theta}(\alpha) = \min\{t : \hat{F}(t) \geq \alpha\}$
- Study design identical to CDF simulation
- Median estimation ($\alpha = 0.5$)

Table 3: MSE ratios for median estimation

Population	h	σ	HT	CD0	CD1	RKM0	RKM1	DORF
Ratio	0.10	0.1	3.14	0.37	0.31	0.97	1.00	1.25
	0.10	0.4	1.15	0.58	1.75	0.89	0.91	0.95
	0.25	0.1	3.09	0.37	0.31	0.96	0.98	1.26
	0.25	0.4	1.26	0.64	1.90	0.97	0.99	1.03
Linear	0.10	0.1	8.14	0.88	0.32	0.85	0.84	1.78
	0.10	0.4	2.49	3.64	0.65	1.05	1.04	1.13
	0.25	0.1	6.55	0.71	0.26	0.69	0.68	1.85
	0.25	0.4	2.57	3.77	0.61	1.08	1.08	1.18
Expo	0.10	0.1	1.32	2.34	7.68	1.80	0.99	1.01
	0.10	0.4	1.00	0.88	0.92	1.13	0.95	0.96
	0.25	0.1	1.40	2.49	8.16	1.92	1.05	1.10
	0.25	0.4	1.06	0.94	0.97	1.21	1.01	1.02
Bump	0.10	0.1	11.70	35.91	26.10	2.53	2.84	1.72
	0.10	0.4	2.66	6.97	2.23	1.26	1.31	1.18
	0.25	0.1	7.01	21.49	15.63	1.51	1.70	1.42
	0.25	0.4	2.37	6.22	1.99	1.12	1.17	1.16
Jump	0.10	0.1	2.01	3.46	3.98	1.46	1.81	1.10
	0.10	0.4	1.18	1.74	0.82	1.07	1.10	0.99
	0.25	0.1	1.88	3.24	3.74	1.37	1.70	1.18
	0.25	0.4	1.26	1.85	0.88	1.14	1.18	1.07
Quad	0.10	0.1	1.98	28.08	1.91	2.66	1.99	1.17
	0.10	0.4	0.96	2.55	0.87	1.25	0.96	0.97
	0.25	0.1	1.98	28.04	1.90	2.65	1.99	1.44
	0.25	0.4	1.02	2.71	0.92	1.33	1.02	1.02
Cycle	0.10	0.1	7.07	29.93	1.04	10.65	2.46	2.06
	0.10	0.4	3.89	18.47	0.86	6.10	1.67	1.30
	0.25	0.1	4.32	18.29	0.63	6.51	1.51	1.83
	0.25	0.4	3.52	16.68	0.78	5.51	1.51	1.55

Table 4: Percent relative biases for median estimation

Population	h	σ	HT	CD0	CD1	RKM0	RKM1	LPR	DORF
Ratio	0.10	0.1	-0.89	-2.06	1.62	-0.64	-0.59	0.47	0.40
	0.10	0.4	-1.16	-3.18	18.15	-1.25	-0.91	0.61	0.92
	0.25	0.1	-0.89	-2.06	1.62	-0.64	-0.59	0.50	0.17
	0.25	0.4	-1.16	-3.18	18.15	-1.25	-0.91	0.32	1.53
Linear	0.10	0.1	-0.61	-2.60	-1.50	-0.27	-0.27	0.72	0.57
	0.10	0.4	-1.67	-10.62	-3.21	-1.68	-1.67	-0.65	-0.66
	0.25	0.1	-0.61	-2.60	-1.50	-0.27	-0.27	0.74	0.48
	0.25	0.4	-1.67	-10.62	-3.21	-1.68	-1.67	-0.67	-0.79
Expo	0.10	0.1	-1.54	-24.88	63.67	-0.95	-1.85	-4.42	-4.29
	0.10	0.4	0.44	-20.88	19.98	1.79	-0.20	-1.58	-3.29
	0.25	0.1	-1.54	-24.88	63.67	-0.95	-1.85	-4.61	-5.71
	0.25	0.4	0.44	-20.88	19.98	1.79	-0.20	-2.79	-8.36
Bump	0.10	0.1	-2.25	-16.61	-14.09	-0.81	-0.89	0.40	0.08
	0.10	0.4	-1.95	-15.32	-8.11	-1.54	-1.44	-0.28	-0.43
	0.25	0.1	-2.25	-16.61	-14.09	-0.81	-0.89	0.48	0.03
	0.25	0.4	-1.95	-15.32	-8.11	-1.54	-1.44	-0.31	-0.86
Jump	0.10	0.1	-0.73	-2.80	-4.22	-0.51	-0.50	-0.26	-0.24
	0.10	0.4	-0.82	-9.18	-1.24	-0.83	-0.49	0.22	0.27
	0.25	0.1	-0.73	-2.80	-4.22	-0.51	-0.50	-0.43	-0.27
	0.25	0.4	-0.82	-9.18	-1.24	-0.83	-0.49	-0.19	0.58
Quad	0.10	0.1	0.01	-0.28	0.17	0.01	-1.11	-0.11	-0.08
	0.10	0.4	0.03	-2.38	0.46	-0.03	-0.05	0.00	-0.22
	0.25	0.1	0.01	-0.28	0.17	0.01	-0.11	-0.05	-0.56
	0.25	0.4	0.03	-2.38	0.46	-0.03	-0.05	-0.05	-0.71
Cycle	0.10	0.1	-1.80	-11.29	-2.25	-1.88	-1.59	-1.96	-1.93
	0.10	0.4	-1.02	-9.26	-1.49	-1.26	-0.54	-0.50	-0.74
	0.25	0.1	-1.80	-11.29	-2.25	-1.88	-1.59	-2.18	-1.98
	0.25	0.4	-1.02	-9.26	-1.49	-1.26	-0.54	-0.74	-0.79

Quantile Simulation Results

- 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
- Area of divergence from CDF simulation results:
 - CD0 estimation for “quad” population
 - LPR has increased efficiency for quantile estimation simulation
 - MSE ratio of 2.71 vs. 0.51

Semiparametric Regression Outline

- Overview
 - handling multiple auxiliary variables
 - motivation for semiparametric methods
- Semiparametric regression
 - Model-assisted survey estimation
 - Application to finite population CDF estimation

Estimation with Multiple Auxiliary Variables

- Multiple auxiliary variables available for entire landscape
- Parametric approach
 - extend CD & RKM to handle all variables
 - but this loses flexibility of nonparametric methodology
- Nonparametric approach
 - continuous case: smoothing in multiple dimensions runs into problems with “curse of dimensionality”
 - categorical case: not meaningful to do kernel smoothing
- Merge nonparametric and parametric methodology

Semiparametric Regression (SEMI)

- Adjust superpopulation to handle additional auxiliary variables: $\mathbf{x}_i = (x_i, \mathbf{z}_i)$

$$\begin{aligned}y_i &= g(x_i, \mathbf{z}_i) + v^{1/2}(x_i)\epsilon_i \\ &= m(x_i) + \mathbf{z}_i\boldsymbol{\beta} + v^{1/2}(x_i)\epsilon_i\end{aligned}$$

where ϵ_i and $v(x_i)$ are as before

- x_i is a single continuous variable
- $\mathbf{z}_i = (1, z_{1i}, z_{2i}, \dots, z_{Di})$ is a vector of D auxiliary variables and allowing for an intercept term
- Model is nonparametric function of x_i plus parametric function of \mathbf{z}_i

General Semiparametric Regression Continued

- Define:

$$\mathbf{\Pi}_A = \text{diag} \{ \pi_i : i \in A \}$$

$$\mathbf{Z}_A = [\mathbf{z}'_i : i \in A]'$$

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

$$\mathbf{S}_A = [\mathbf{s}'_{Ai} : i \in A]$$

- Centered sample smoother matrix ($n \times n$):

$$\mathbf{S}_A^* = (\mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{\Pi}_A^{-1}/N)\mathbf{S}_A$$

- Design-weighted estimator of $\boldsymbol{\beta}$:

$$\hat{\mathbf{B}} = (\mathbf{Z}'_A\mathbf{\Pi}_A^{-1}(\mathbf{I} - \mathbf{S}_A^*)\mathbf{Z}_A)^{-1}\mathbf{Z}'_A\mathbf{\Pi}_A^{-1}(\mathbf{I} - \mathbf{S}_A^*)\mathbf{y}_A$$

- Design-weighted estimator of $m(x_i)$:

$$\hat{m}_i = \mathbf{s}'_{Ai}(\mathbf{y}_A - \mathbf{Z}_A\hat{\mathbf{B}})$$

SEMI in Survey Sampling Estimation

- Breidt and Opsomer (working paper) SEMI estimator for population total:

$$\hat{T}_{SEMI} = \sum_{i \in U} \hat{g}_i + \sum_{i \in A} \frac{y_i - \hat{g}_i}{\pi_i}$$

where $\hat{g}_i = \hat{g}(x_i, \mathbf{z}_i) = \hat{m}_i + \mathbf{z}_i \hat{\mathbf{B}}$

- Design properties of \hat{T}_{SEMI} :
 - semiparametric, model-assisted
 - design consistent
 - asymptotically design unbiased
- Estimated variance of \hat{T}_{SEMI} :

$$\widehat{\text{Var}}(\hat{T}_{SEMI}) = \sum_{i, j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i - \hat{g}_i}{\pi_i} \frac{y_j - \hat{g}_j}{\pi_j}$$

Semiparametric CDF Estimator

- Based on \hat{T}_{SEMI}
- Replace y_i in \hat{T}_{SEMI} with $I_{\{y_i \leq t\}}$:

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

where

$$\hat{\mathbf{B}} = (\mathbf{Z}'_A \mathbf{\Pi}_A^{-1} (\mathbf{I} - \mathbf{S}_A^*) \mathbf{Z}_A)^{-1} \mathbf{Z}'_A \mathbf{\Pi}_A^{-1} (\mathbf{I} - \mathbf{S}_A^*) \mathbf{I}_A$$

$$\hat{m}_i = \mathbf{s}'_{Ai} (\mathbf{I}_A - \mathbf{Z}_A \hat{\mathbf{B}})$$

$$\hat{g}_i = \hat{m}_i + \mathbf{z}_i \hat{\mathbf{B}}$$

Semiparametric CDF Estimator Continued

- Design properties of \hat{T}_{SEMI} hold for \hat{F}_{SEMI} :
 - semiparametric, model-assisted
 - design consistent
 - asymptotically design unbiased
- Estimated variance of $\hat{F}_{SEMI}(t)$:

$$\widehat{\text{Var}}(\hat{F}_{SEMI}(t)) = \frac{1}{N^2} \sum_{i,j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{I_{ti} - \hat{g}_i}{\pi_i} \frac{I_{tj} - \hat{g}_j}{\pi_j}$$

Empirical Example Outline

- Acidity of Northeastern lakes
- Study design
 - EMAP survey of Northeastern lakes
 - Estimating percentage of acidic lakes
- Numerical results

Acidity of Northeastern Lakes

- Acid sensitivity of Northeastern lakes
 - National Surface Water Survey (NSWS)
 - * 1984–1986
 - * 4.2 percent of Northeastern lakes acidic
 - CAAA (1990) placed restrictions on industrial sulfur and nitrogen emissions
- Acid neutralizing capacity (ANC)
 - water's ability to buffer acid
 - $ANC < 0$ indicates the presence of acidity
- What effect have CAAA restrictions had on acidity of these lakes?

EMAP Survey of Northeastern lakes

- 1991 through 1996
- $N = 21384$ lakes, 557 water samples
 - some lakes sampled multiple times
 - average multiple measurements to obtain 1 measurement per sampled lake
 - $n = 338$
- Treat as stratified sample with replacement
- ANC only available for sample
- Multiple auxiliary variables for entire landscape
- How do we determine the proportion of lakes with $ANC < 0$?

CDF Estimation for ANC

- Problem: want to estimate $F(0)$ for $y =$ ANC

- Auxiliary variables:

x_i = longitude

z_{ji} = indicator of eco-region j

z_{11i} = latitude

z_{12i} = elevation

– $j = 1, \dots, 10$

– eco-region is categorical

– covariate space includes empty holes

- Estimators:

– design-based Hájek:

$$\hat{F}_{HT*}(t) = \left(\sum_{j \in A} \frac{1}{\pi_j} \right)^{-1} \sum_{i \in A} \frac{I_{\{y_i \leq t\}}}{\pi_i}$$

– SEMI (\hat{F}_{SEMI}):

model is nonparametric in x_i and

parametric in $\mathbf{z}_i = (1, z_{1i}, z_{2i}, \dots, z_{12i})$

Empirical Results

	$\hat{F}(0)$	$\text{StDev}(\hat{F}(0))$
SEMI	0.044	0.0176
HT*	0.059	0.0214

95% CI for $F(0)$ based on $\hat{F}_{SEMI}(0)$:

$$(0.044 \pm 1.96(0.0176)) = (0.0095, 0.0785)$$

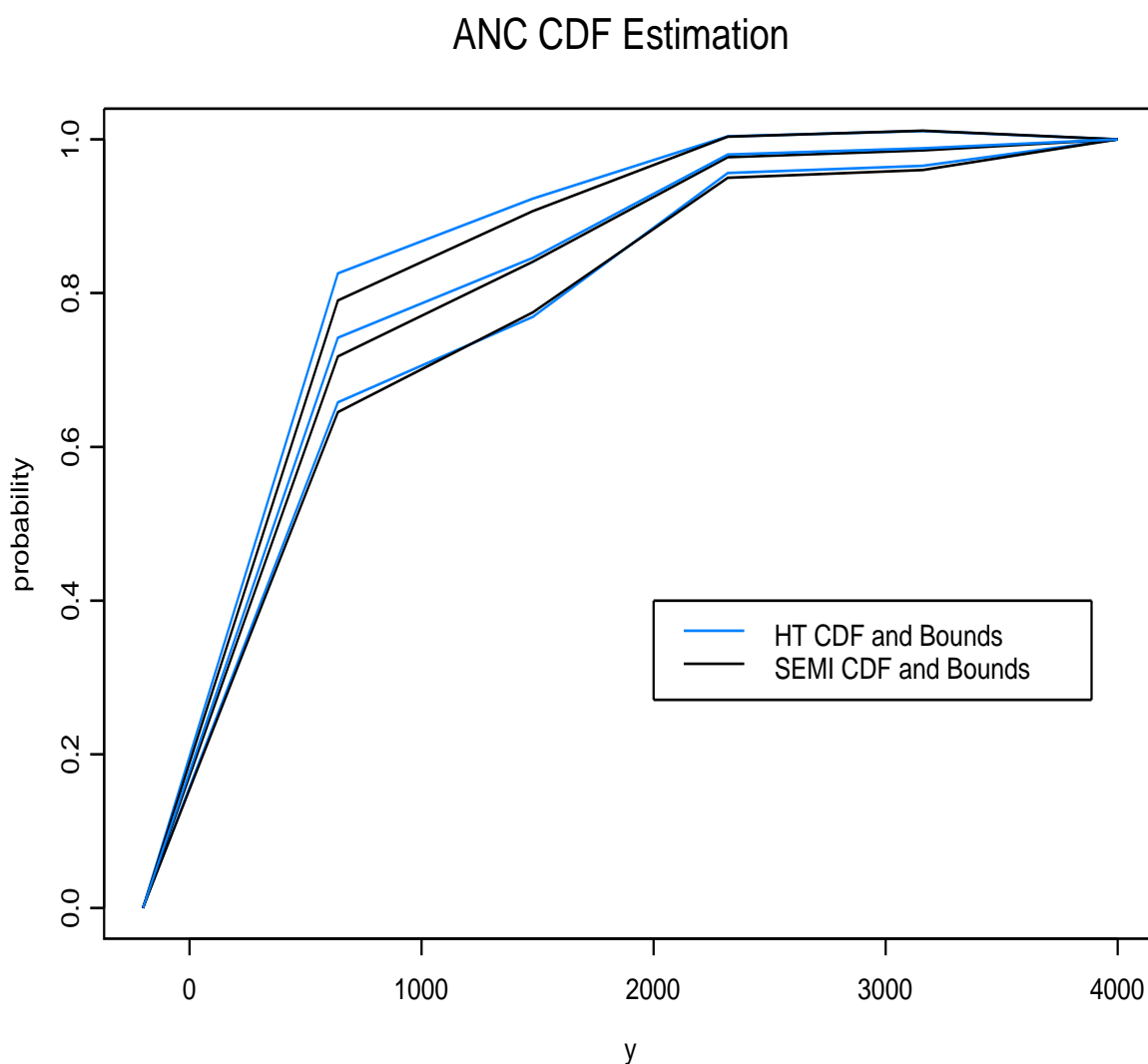
95% CI for $F(0)$ based on $\hat{F}_{HT*}(0)$:

$$(0.059 \pm 1.96(0.0214)) = (0.0170, 0.1010)$$

- CI based on $\hat{F}_{HT*}(0)$ is 22 percent wider
- NSW estimate of 4.2% is within both CI's
- No evidence that CAAA emissions restrictions have improved or worsened levels of acidity

ANC CDF Estimation

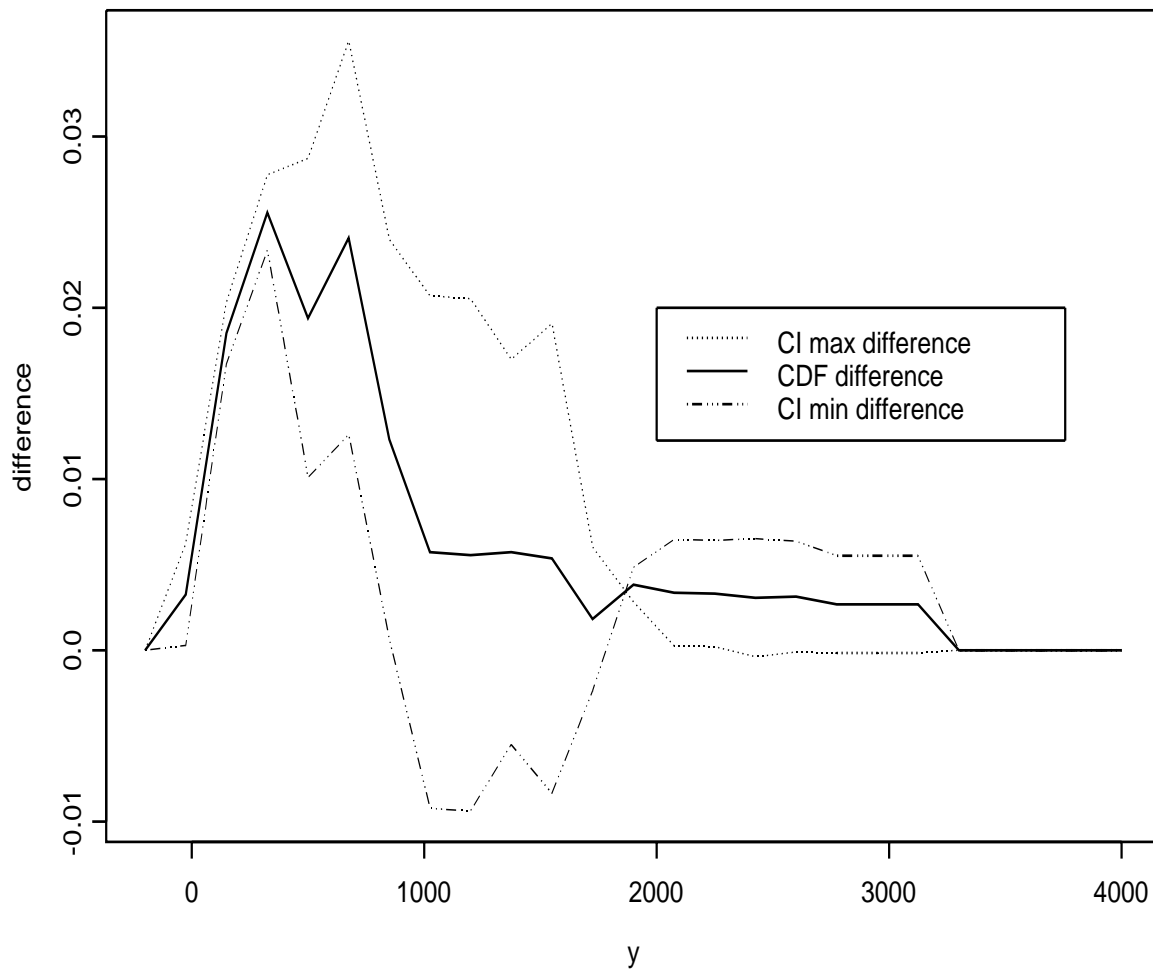
- Evaluate $\hat{F}_{HT*}(t)$ and $\hat{F}_{SEMI}(t)$ for 1000 grid points in the range of ANC



- CI's based on $\hat{F}_{HT*}(t)$ are 9 percent wider on average

CI Comparison

CI Comparison for ANC CDF Estimation



$$\begin{aligned}\text{CI max difference} &= \hat{F}_{HT^*}(t) \text{ upper bound} - \hat{F}_{SEMI}(t) \text{ upper bound} \\ \text{CI min difference} &= \hat{F}_{HT^*}(t) \text{ lower bound} - \hat{F}_{SEMI}(t) \text{ lower bound} \\ \text{CDF difference} &= \hat{F}_{HT^*}(t) - \hat{F}_{SEMI}(t)\end{aligned}$$

Extensions

- SEMI easily handles additional auxiliary variables
- SEMI easily extended to other survey estimates and study variables
- Quantile estimation is straightforward:

$$\hat{\theta}_{SEMI}(\alpha) = \min\{t : \hat{F}_{SEMI}(t) \geq \alpha\}$$

Quantile Estimates of Chemistry Variables

α	Sulfate	Magnesium	Chloride
0.25	73.0	66.1	25.1
0.50	105.0	123.8	177.0
0.75	194.9	238.3	495.4

Conclusion

- CDF estimation with one auxiliary variable
 - Local polynomial regression CDF estimator
 - Monte Carlo comparison of several estimators

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

- CDF estimation with multiple auxiliary variables
 - semiparametric CDF estimator
 - Northeastern lakes example

References

- Breidt, F.J. and J.D. Opsomer (2000). Local polynomial regression estimators in survey sampling. *Ann. Statist.*, **28**, 1026–1053.
- Breidt, F.J. and J.D. Opsomer (2002). Design Properties of Semiparametric Model-assisted Estimators. Working paper. Iowa State University.
- Chambers, R.L., A.H. Dorfman, and P. Hall (1992). Properties of estimators of the finite population distribution function. *Biometrika* **79**, 577–82.
- Chambers, R.L. and R. Dunstan (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.
- Larsen, D.P., K.W. Thornton, N.S. Urquhart, and S.G. Paulsen (1993). Overview of Survey Design and Lake Selection. *EMAP - Surface Waters 1991 Pilot Report*, edited by Larsen, D.P. and Christie, S.J. EPA/620/R-93/003.
- Rao, J.N.K., J.G. Kovar, and H.J. Mantel (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* **77**, 365–75.
- Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Stoddard, J.L., J.S. Kahl, F.A. Deviney, D.R. DeWalle, C.T. Driscoll, A.T. Herlihy, J.H. Kellogg, P.S. Murdoch, J.R. Webb, and K.E. Webster (2002). Response of surface water chemistry to the Clean Air Act Amendments of 1990.

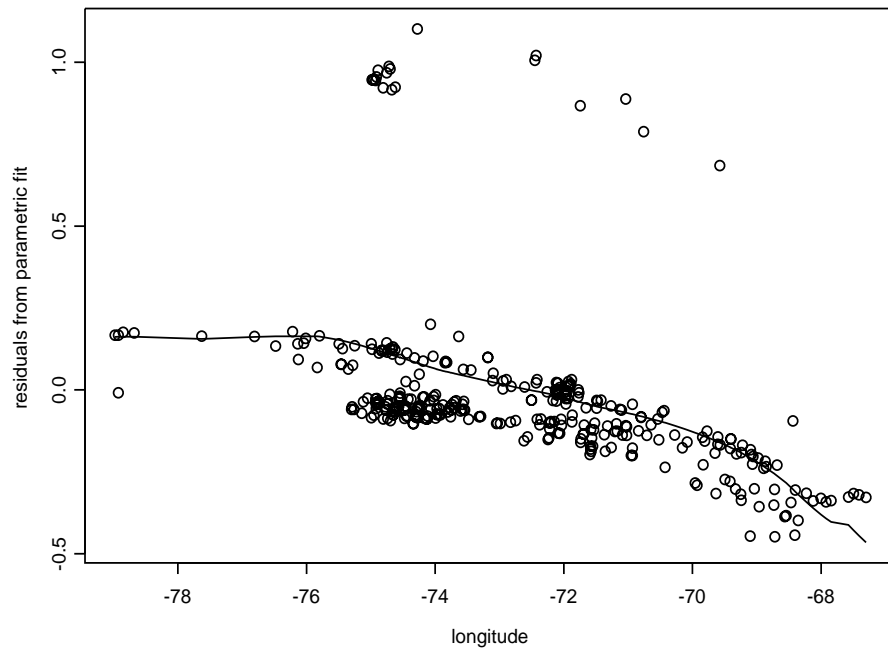
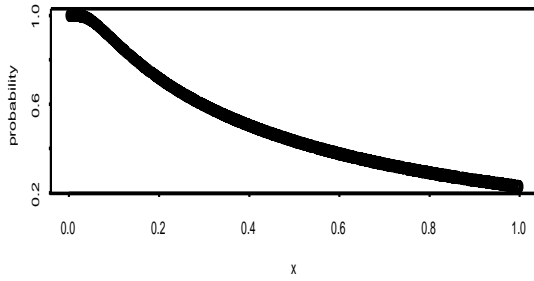
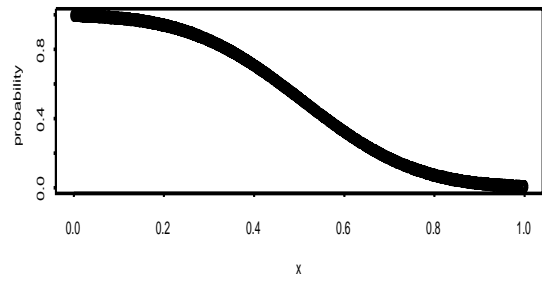


Figure 1: Residuals from a parametric fit versus longitude with superimposed nonparametric smooth based on a bandwidth equal to 5

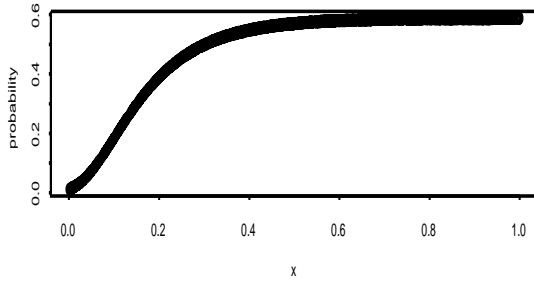
ratio, t=median, sig=0.4



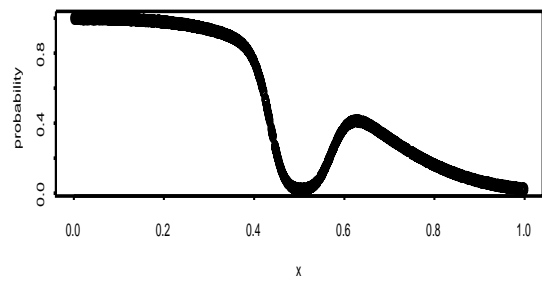
linear, t=median, sig=0.4



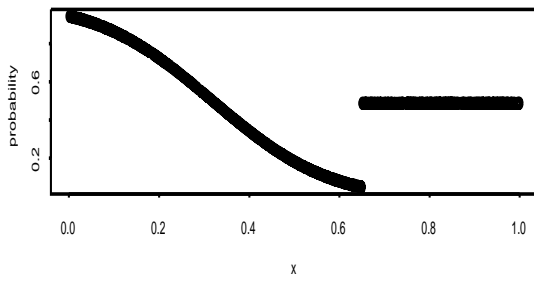
expo, t=median, sig=0.4



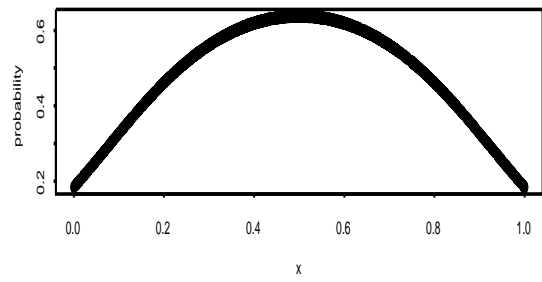
bump, t=median, sig=0.4



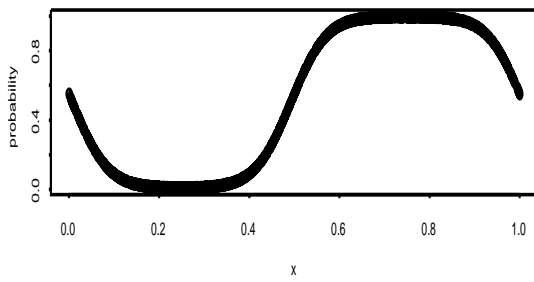
jump, t=median, sig=0.4



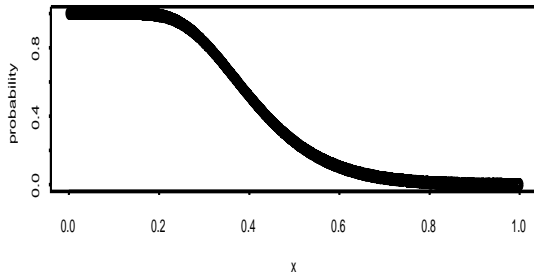
quad, t=median, sig=0.4



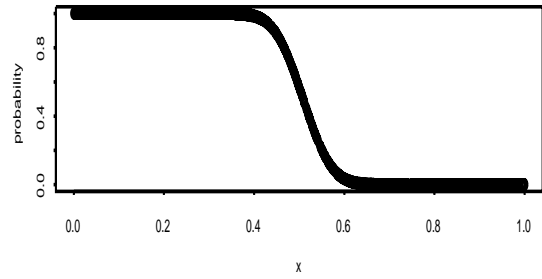
cycle, t=median, sig=0.4



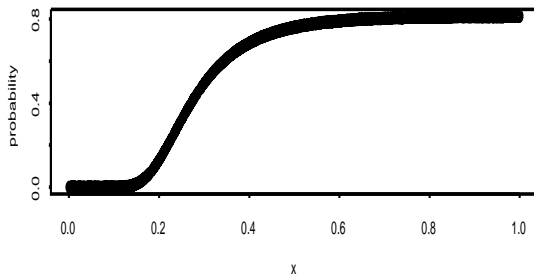
ratio, t=median, sig=0.1



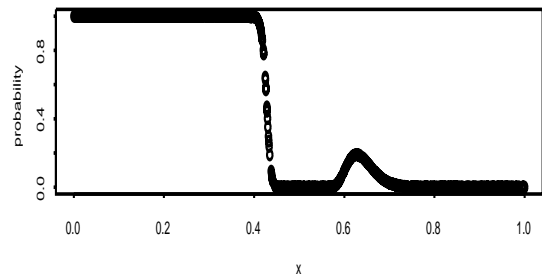
linear, t=median, sig=0.1



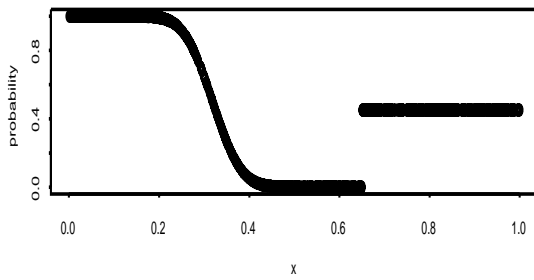
expo, t=median, sig=0.1



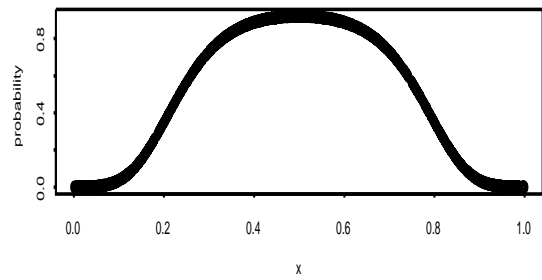
bump, t=median, sig=0.1



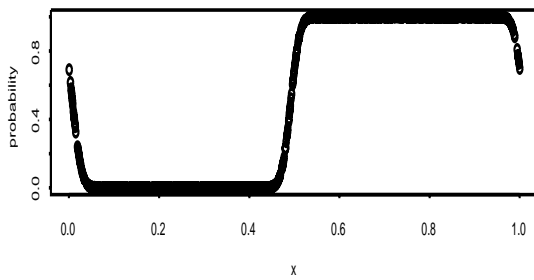
jump, t=median, sig=0.1



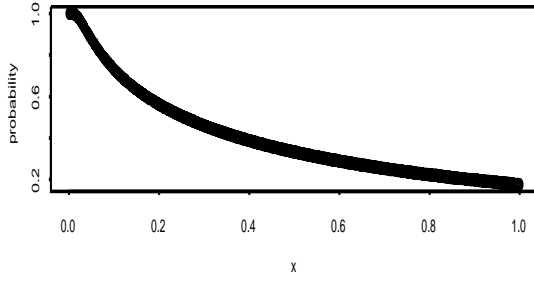
quad, t=median, sig=0.1



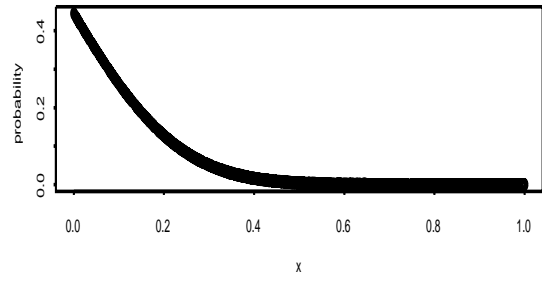
cycle, t=median, sig=0.1



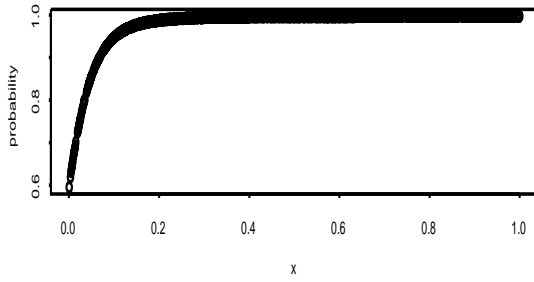
ratio, t=first, sig=0.4



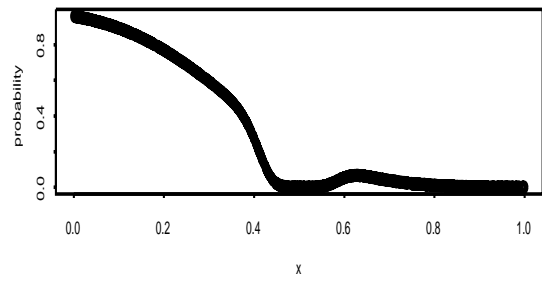
linear, t=first, sig=0.4



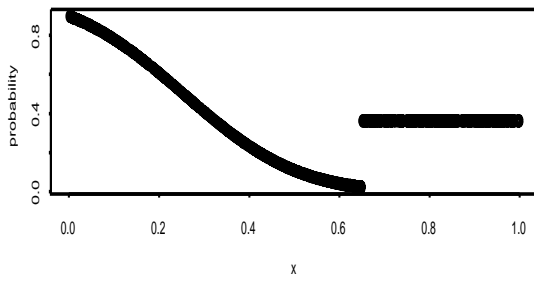
expo, t=first, sig=0.4



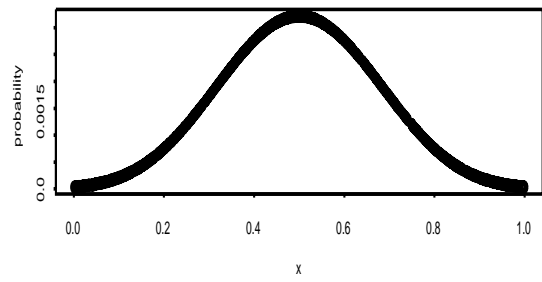
bump, t=first, sig=0.4



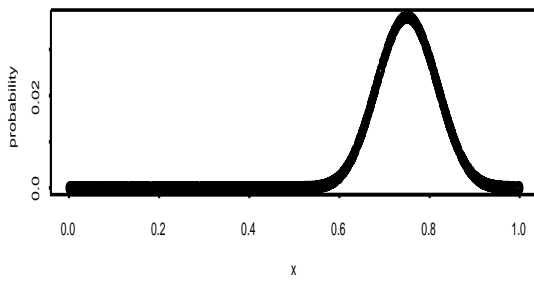
jump, t=first, sig=0.4



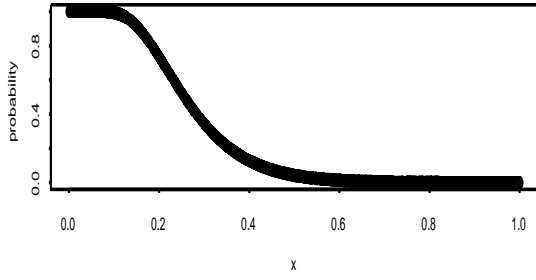
quad, t=first, sig=0.4



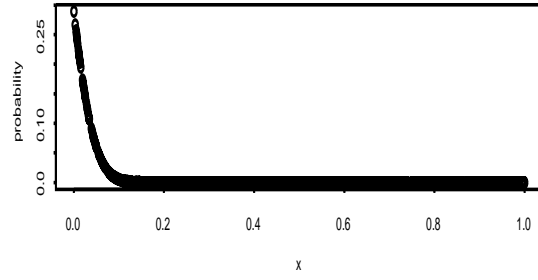
cycle, t=first, sig=0.4



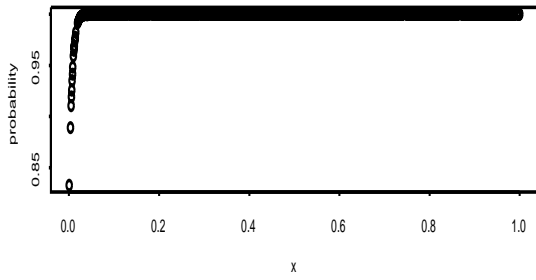
ratio, t=first, sig=0.1



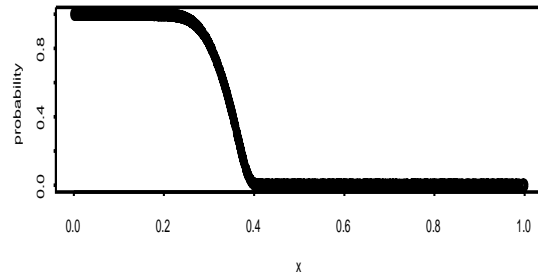
linear, t=first, sig=0.1



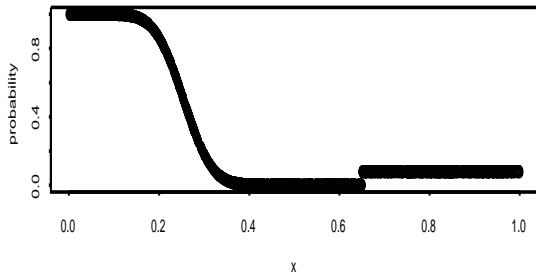
expo, t=first, sig=0.1



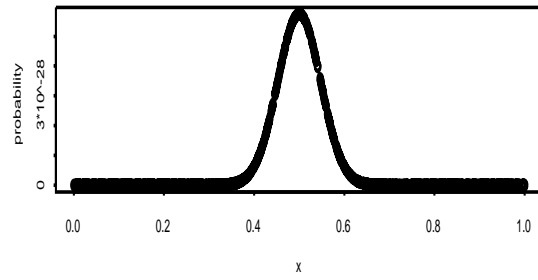
bump, t=first, sig=0.1



jump, t=first, sig=0.1



quad, t=first, sig=0.1



cycle, t=first, sig=0.1

