

Semiparametric Model-Assisted Estimation of Distribution Functions in Surveys with Auxiliary Information

Alicia Johnson

Department of Statistics
Colorado State University

Joint work with F. Jay Breidt, CSU
and Jean Opsomer, Iowa State University

The work reported here was developed under the 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 authors. EPA does not endorse any products or commercial services mentioned in this report.

Outline

- Introduction
 - finite population CDF estimation for y
 - design-based Hájek estimator
- Estimation with auxiliary information \mathbf{x}
 - auxiliary information \mathbf{x} available for entire landscape
 - existing methods that incorporate \mathbf{x}
- Estimation with multiple auxiliary variables
 - motivation for semiparametric methods
- Semiparametric regression
 - model-assisted survey estimation
 - extension to cdf estimation
- Empirical results
 - acidity of Northeastern lakes
 - comparison of SEMI and Hájek

Finite Population CDF Estimation

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

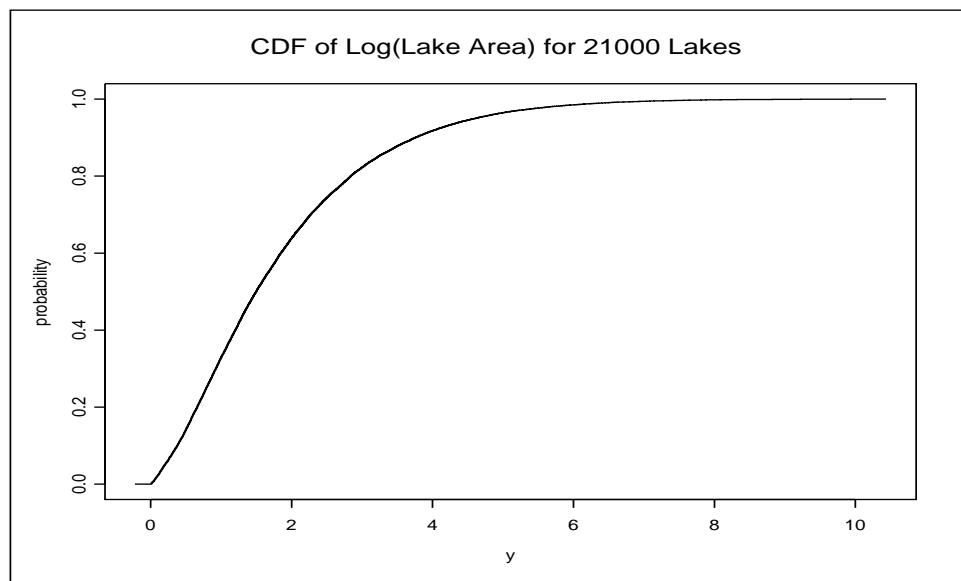
- Some Notation:

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

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

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

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



Hájek Estimator

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

- Note: $\left(\sum_{j \in s} \frac{1}{\pi_j} \right)^{-1} = \frac{1}{N}$ for SI design
- Asymptotically design unbiased
(exactly unbiased for equal probability)
- No dependence on any model
- Does not incorporate auxiliary information \mathbf{x}
- How do we incorporate \mathbf{x} for the entire universe?

CDF Estimation with Auxiliary Information

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

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

- Parametric methods
 - model-based: Chambers & Dunstan (1986)
 - model-assisted: Rao, Kovar, & Mantel (1990)
- Nonparametric methods
 - model-based: Dorfman (1992) [totals]
 - model-assisted: Breidt & Opsomer (2000)
Johnson (2003) [CDF]

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: $\mathbf{x}_i = (x_i, \mathbf{a}_i)$

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

where ϵ_i and $v(x_i)$ are as before

- x_i is a single continuous variable
- $\mathbf{a}_i = (1, a_{1i}, a_{2i}, \dots, a_{Di})$ is a vector of $D+1$ auxiliary variables
- Model is nonparametric function of x_i plus parametric function of \mathbf{a}_i

Semiparametric Regression Estimator

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

$$\hat{T}_{SEMI} = \underbrace{\sum_{i \in U} \hat{g}_i}_{\text{model-based prediction}} + \underbrace{\sum_{i \in s} \frac{y_i - \hat{g}_i}{\pi_i}}_{\text{design bias adjustment}}$$

where:

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

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

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

Semiparametric Regression Estimator Continued

- Replace y_i 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:

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

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

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

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

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

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 $y = \text{ANC}$
- Auxiliary variables:
 - x_i = longitude
 - a_{ji} = indicator of eco-region j
 - $a_{11,i}$ = latitude
 - $a_{12,i}$ = elevation
- $j = 1, \dots, 10$
- eco-region is categorical
- covariate space includes empty holes
- Estimators:
 - design-based Hájek
 - SEMI (\hat{F}_{SEMI}):
 - model is nonparametric in x_i and
 - parametric in $\mathbf{a}_i = (1, a_{1i}, a_{2i}, \dots, a_{12i})$

Empirical Results

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

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

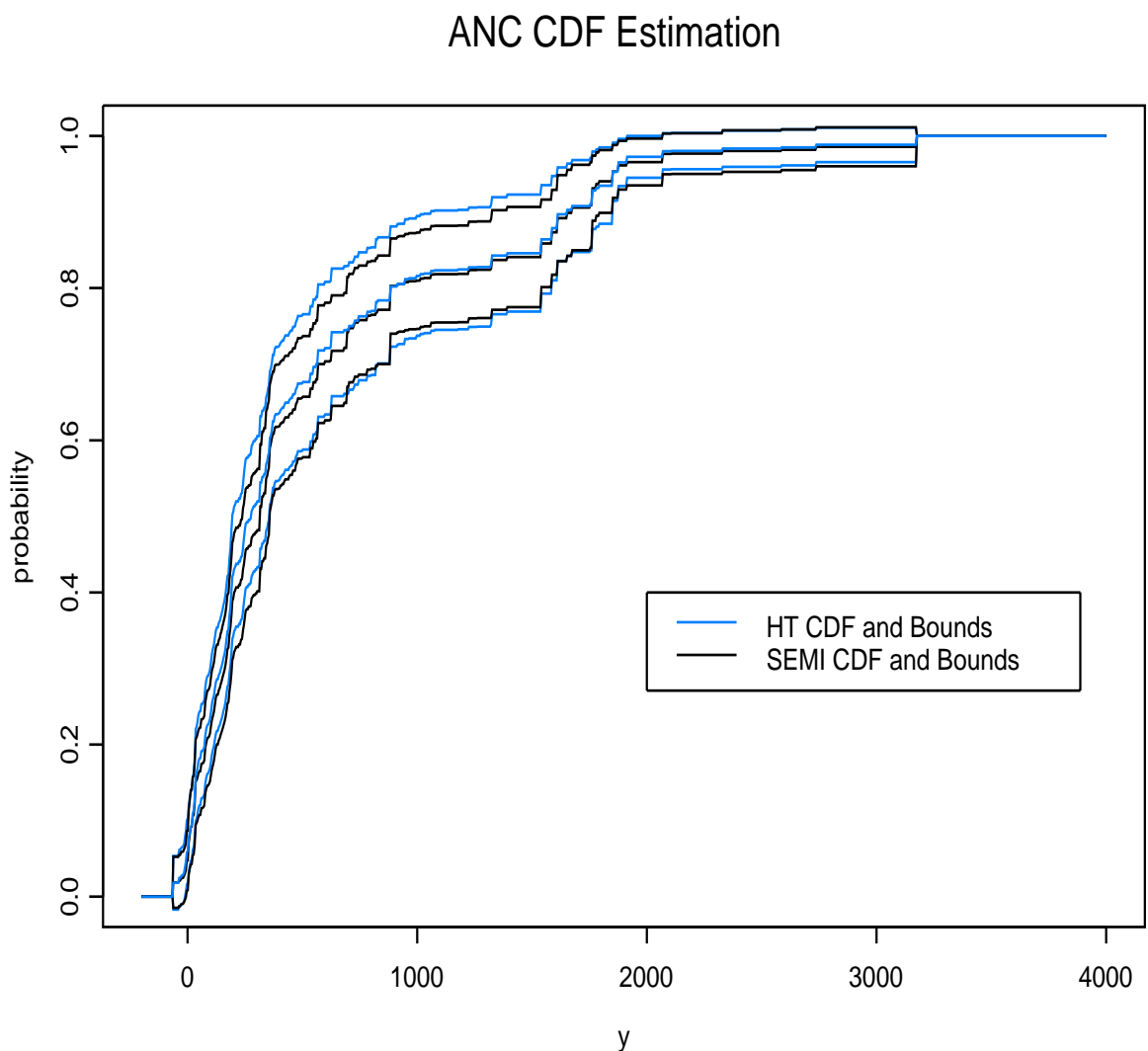
95% CI for $F(0)$ based on $\hat{F}_H(0)$:

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

- CI based on $\hat{F}_H(0)$ 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

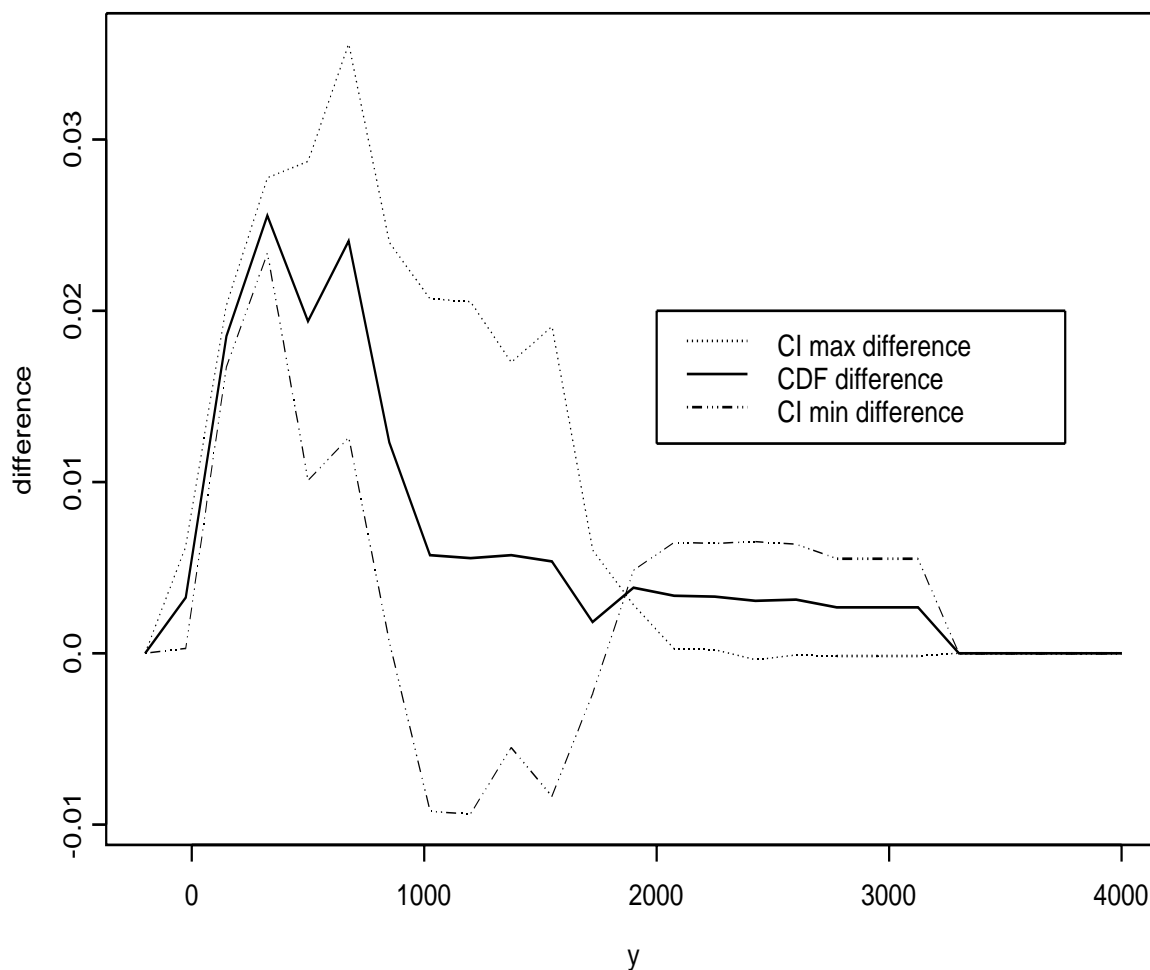
- Evaluate $\hat{F}_H(t)$ and $\hat{F}_{SEMI}(t)$ for 1000 grid points in the range of ANC



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

Numerical Results Continued

CI Comparison for ANC CDF Estimation



$$\text{CI max difference} = \hat{F}_H(t) \text{ upper bound} - \hat{F}_{SEMI}(t) \text{ upper bound}$$

$$\text{CI min difference} = \hat{F}_H(t) \text{ lower bound} - \hat{F}_{SEMI}(t) \text{ lower bound}$$

$$\text{CDF difference} = \hat{F}_H(t) - \hat{F}_{SEMI}(t)$$

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

Summary

- CDF Estimation with multiple auxiliary variables
- Semiparametric CDF estimator
 - model-assisted
 - design consistent
 - asymptotically design unbiased
- Empirical example
 - acidity of Northeastern lakes
 - SEMI vs. Hájek

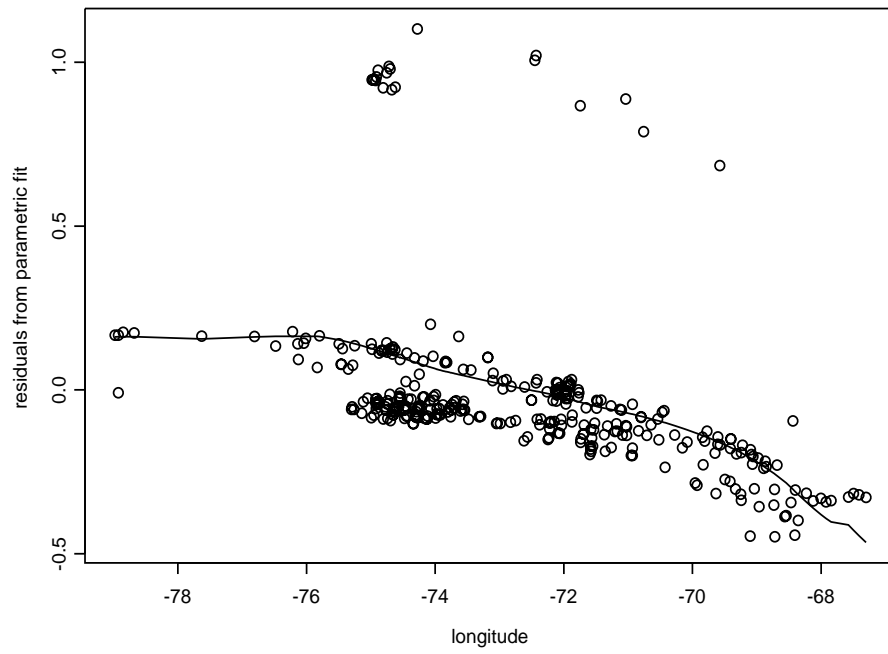


Figure 1: Residuals from a parametric fit versus longitude with superimposed nonparametric smooth based on a bandwidth equal to 5

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 Semi-parametric 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.