

Local Polynomial Regression Estimators in Survey Sampling

F. Jay Breidt

Joint work with Jean D. Opsomer

Iowa State University

June 28, 1999

Outline

- Use of auxiliary information in surveys
 - review of estimation procedures
 - operational considerations
- Local polynomial regression estimators
 - asymptotic properties
 - simulation results
- Example
 - soil mapping
- Summary
 - future work

Auxiliary Information

- Finite population $U_N = \{1, 2, \dots, N\}$
- Draw sample $s \subset U_N$ (sample size n_N) via $p(\cdot)$
- Observe study variables, $y_i, i \in s$
- Obtain complete auxiliary information $x_i, i \in U_N$
 - household-specific rent or housing value on U.S. Census
 - individual-specific taxable income on Swedish population register
 - pixel-specific spectral value from Landsat image
- Modeling?

Operational Aspects

- Often, a statistical agency:
 - collects data and auxiliary info
 - releases basic tabular estimates and weighted data set
- Modeling constraints:
 - limited time and other resources for modeling y_i 's
 - potential controversy among end users
 - confidentiality restrictions on x_i
- Estimation strategy:
 - should use information in $x_i, i \in U_N$
 - should *not* release $x_i, i \in U_N$
 - should handle any study variables (and be internally consistent)
 - should *not* require modeling efforts for every study variable
 - should be efficient if model is right
 - should *not* fail if model is wrong

Generalized Difference Estimator

- For *fixed* y_i^0 ,

$$\hat{t}_{\text{diff}} = \sum_{i \in U_N} y_i^0 + \sum_{i \in U_N} (y_i - y_i^0) \frac{I_i}{\pi_i}$$

– Horvitz-Thompson is special case with $y_i^0 \equiv 0$

- Design expectation is

$$E_p [\hat{t}_{\text{diff}}] = \sum_{i \in U_N} y_i^0 + \sum_{i \in U_N} (y_i - y_i^0) = t_y$$

- Design variance is

$$\text{Var}_p (\hat{t}_{\text{diff}}) = \sum_{i,j \in U_N} (y_i - y_i^0)(y_j - y_j^0) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} = O\left(\frac{N^2}{n_N}\right)$$

- Unbiased variance estimator is

$$\hat{V}(\hat{t}_{\text{diff}}) = \sum_{i,j \in U_N} (y_i - y_i^0)(y_j - y_j^0) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{I_i I_j}{\pi_{ij}}$$

Model-Assisted Approach

- Superpopulation model, ξ : y_i 's are independent r.v.'s with model mean

$$E_{\xi} [y_i] = m(x_i)$$

and model variance

$$\text{Var}_{\xi} (y_i) = v(x_i)$$

- Estimate parameters of model ξ and get model-based predictions \hat{m}_i
- Replace y_i^0 in GDE with \hat{m}_i :

$$\hat{t}_y = \sum_{i \in U_N} \hat{m}_i + \sum_{i \in U_N} (y_i - \hat{m}_i) \frac{I_i}{\pi_i}$$

- model-based prediction + design bias adjustment
- if model is good, should have small design variance
- With $E_{\xi} [y_i] = \mathbf{x}'_i \boldsymbol{\beta}$, get generalized regression (GREG) estimator
 - ADU and design-consistent (Robinson and Särndal, 1982)

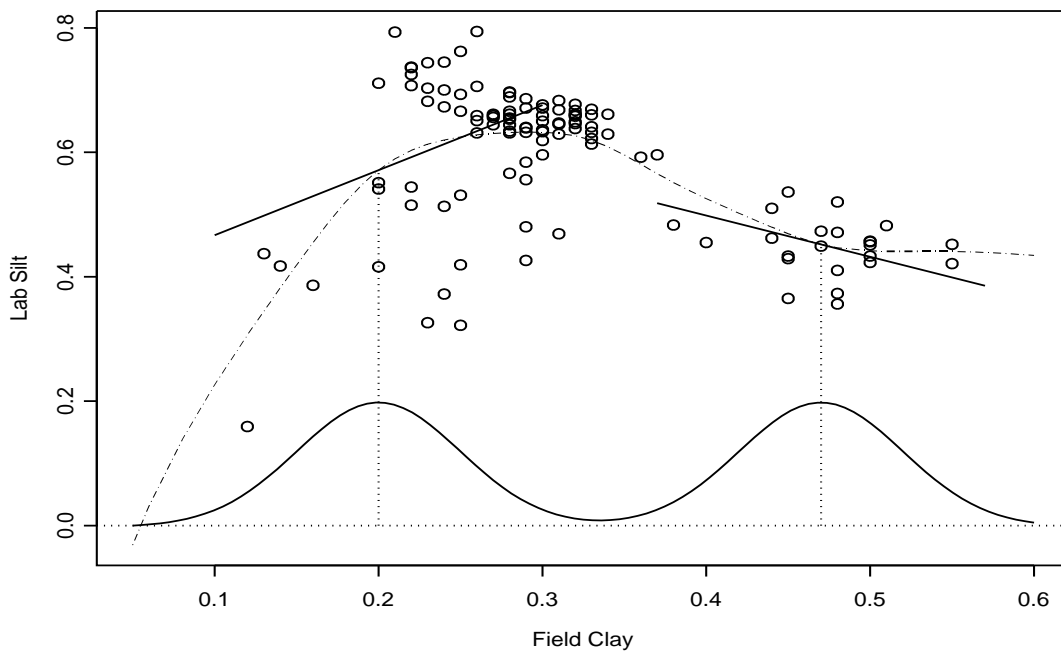
Generalized Regression Estimator: Examples

- Regression through origin with $v(x_i) \propto x_i$: classic ratio estimator
- Heteroskedastic ANOVA: poststratification estimator
- Simple linear regression with $m(x_i) = \beta_0 + \beta_1 x_i$, $v(x_i) = \sigma^2$: classic regression estimator

$$\begin{aligned}\hat{t}_{\text{greg}} &= \sum_{i \in U_N} \mathbf{x}'_i \hat{\mathbf{B}} + \sum_{i \in U_N} (y_i - \mathbf{x}'_i \hat{\mathbf{B}}) \frac{I_i}{\pi_i} \\ &= N \left\{ \bar{y}_s + \hat{B}_1 (\bar{x}_U - \bar{x}_s) \right\}\end{aligned}$$

Local Polynomial Regression

- Nonparametric model, ξ : $m(x_i)$ is smooth, $v(x_i)$ is positive and smooth
 - Kuo (1988), Dorfman (1992), Dorfman and Hall (1993)
- Locally weighted least squares fits (Wand and Jones, 1995)



Local Polynomial Regression Estimator

- Model: $m(x_i)$ is smooth, $v(x_i)$ is positive and smooth
- Define

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

and

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

- Local polynomial regression estimator of $m(x_i)$ based on s is

$$\hat{m}_i = [1, 0, \dots, 0] (\mathbf{X}'_{si} \mathbf{W}_{si} \mathbf{X}_{si})^{-1} \mathbf{X}'_{si} \mathbf{W}_{si} \mathbf{y}_s$$

- LPR estimator of t_y is then

$$\hat{t}_{\text{lpr}} = \sum_{i \in U_N} \hat{m}_i + \sum_{i \in U_N} (y_i - \hat{m}_i) \frac{I_i}{\pi_i}$$

Weighting and Calibration

- Define

$$\mathbf{w}_{si} = [1, 0, \dots, 0] (\mathbf{X}'_{si} \mathbf{W}_{si} \mathbf{X}_{si})^{-1} \mathbf{X}'_{si} \mathbf{W}_{si}$$

- Weights

$$\begin{aligned} \hat{t}_{\text{lpr}} &= \sum_{i \in s} \left\{ \frac{1}{\pi_i} + \sum_{j \in U_N} \left(1 - \frac{I_j}{\pi_j} \right) \mathbf{w}'_{sj} \mathbf{e}_i \right\} y_i \\ &= \sum_{i \in s} \omega_{is} y_i \end{aligned}$$

where \mathbf{e}_i is the i th column of the $n_N \times n_N$ identity matrix

- internally consistent:

$$\hat{t}_{y+z} = \sum_{i \in s} \omega_{is} (y_i + z_i) = \sum_{i \in s} \omega_{is} y_i + \sum_{i \in s} \omega_{is} z_i = \hat{t}_y + \hat{t}_z$$

- can be applied to any study variable

- Calibration (Deville and Särndal, 1992)

$$\sum_{i \in s} \omega_{is} x_i^\ell = \sum_{i \in U_N} x_i^\ell \quad (\ell = 0, 1, \dots, q)$$

- if $y_i = \beta_0 + \beta_1 x_i + \dots + \beta_q x_i^q$, then $\hat{t}_{\text{lpr}} = t_y$ for every possible sample

Asymptotic Framework

- Population assumptions:

- population size $N \rightarrow \infty$, $n_N N^{-1} \rightarrow \pi \in (0, 1)$
- $m(x)$ smooth, $v(x)$ smooth and positive
- noise $\epsilon_i := y_i - m(x_i) \sim (0, v(x_i))$
- regressors $\{x_i\}$ behave like iid sample

- Smoothing assumptions:

- kernel $K(\cdot)$ is symmetric, continuous, and compactly supported
- bandwidth $h_N \rightarrow 0$ and $Nh_N^2 \rightarrow \infty$

- Design assumptions:

- $\min_{i \in U_N} \pi_i \geq \lambda > 0$ and $\min_{i,j \in U_N} \pi_{ij} \geq \lambda^* > 0$
- limited dependence:

$$\limsup_{N \rightarrow \infty} n_N \max_{i,j \in U_N: i \neq j} |\pi_{ij} - \pi_i \pi_j| < \infty$$
$$\limsup_{N \rightarrow \infty} N^2 \max_{(i,j,k,\ell) \text{ distinct}} |\mathbb{E}_p [(I_i - \pi_i)(I_j - \pi_j)(I_k - \pi_k)(I_\ell - \pi_\ell)]| < \infty$$

...

Asymptotic Properties of LPR Estimator

- Asymptotically design unbiased (ADU):

$$\lim_{N \rightarrow \infty} \mathbb{E}_p \left[\frac{\hat{t}_{\text{lpr}} - t_y}{N} \right] = 0 \text{ with model probability one}$$

- Design consistent:

$$\lim_{N \rightarrow \infty} \mathbb{E}_p \left[\mathbf{1}_{\left\{ N^{-1} |\hat{t}_{\text{lpr}} - t_y| > \eta \right\}} \right] = 0 \text{ with model probability one}$$

for all $\eta > 0$

- Design mean squared error:

$$\mathbb{E}_p \left(\frac{\hat{t}_{\text{lpr}} - t_y}{N} \right)^2 = \frac{1}{N^2} \sum_{i,j \in U_N} (y_i - m_i)(y_j - m_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} + o(n_N^{-1})$$

with consistent, ADU estimator

$$\hat{V}(N^{-1} \hat{t}_{\text{lpr}}) = \frac{1}{N^2} \sum_{i,j \in U_N} (y_i - \hat{m}_i)(y_j - \hat{m}_j) \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{I_i I_j}{\pi_{ij}}$$

Godambe-Joshi Lower Bound

- Godambe and Joshi (1965): for any estimator \hat{t}_y satisfying

$$E_p \left[N^{-1}(\hat{t}_y - t_y) \right] = 0,$$

the following inequality holds:

$$E_\xi E_p \left(\frac{\hat{t}_y - t_y}{N} \right)^2 \geq \frac{1}{N^2} \sum_{i \in U_N} v(x_i) \frac{1 - \pi_i}{\pi_i}$$

- LHS is *anticipated variance*
- RHS minimized for $\pi_i \propto v^{1/2}(x_i)$

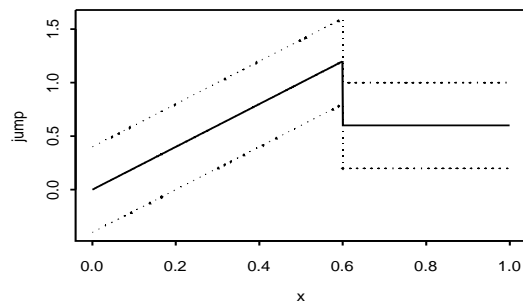
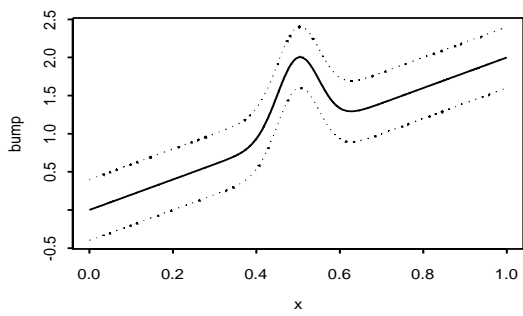
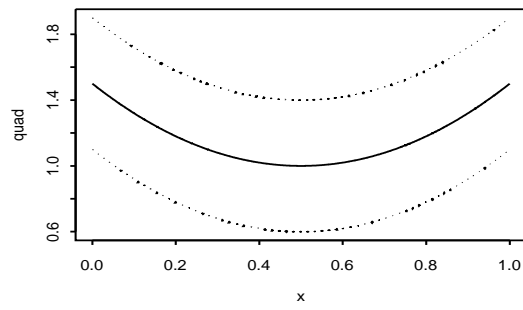
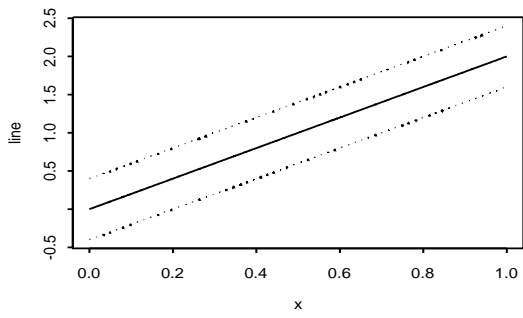
- LPR asymptotically attains GJLB:

$$E_\xi E_p \left(\frac{\hat{t}_{\text{lpr}} - t_y}{N} \right)^2 = \frac{1}{N^2} \sum_{i \in U_N} v(x_i) \frac{1 - \pi_i}{\pi_i} + o(n_N^{-1})$$

- GREG attains GJLB if model is linear (Wright, 1983)

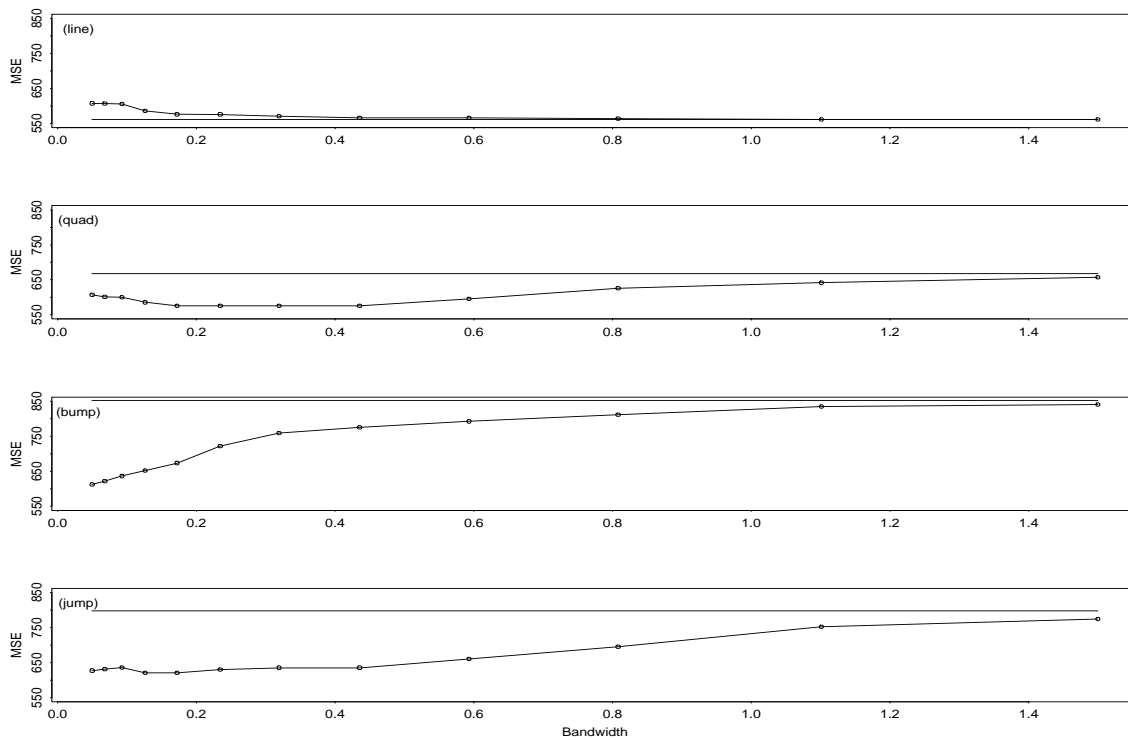
Simulation Experiment

- $N = 1000$, $\{\epsilon_i\}$ iid $N(0, \sigma^2)$



Bandwidth Selection

- Epanechnikov kernel, $0.75(1 - t^2)_+$
- $n_N = 200$, $\sigma = 0.4$



Simulation Results

For $h_N = 0.25$ and 100 simple random samples of size $n_N = 100$,

- Absolute relative design bias, $|\mathbb{E}_p [\hat{t} - t_y] / t_y| \times 100\% < 1\%$
- Relative efficiency, $\text{Var}_p(\hat{t}) / \text{Var}_p(\hat{t}_{\text{1pr}})$:

	line	quad	bump	jump
HT	28.67	2.52	9.67	3.97
REG	0.96	2.55	1.86	3.42

- Mean of estimated standard error over MC standard error:

	line	quad	bump	jump
HT	1.12	1.14	1.13	1.10
REG	1.02	1.12	1.18	1.07
LLR	0.99	1.00	1.19	1.09

Soil Mapping

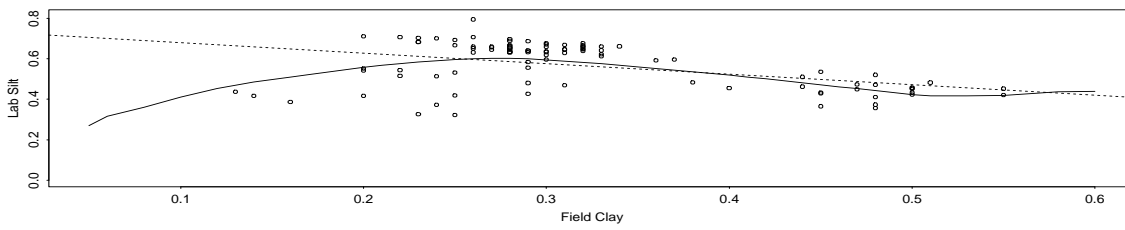
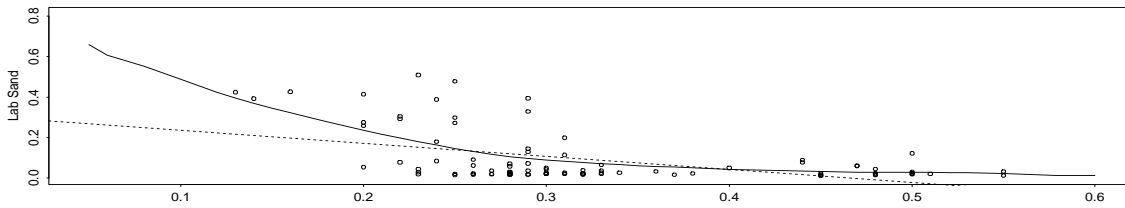
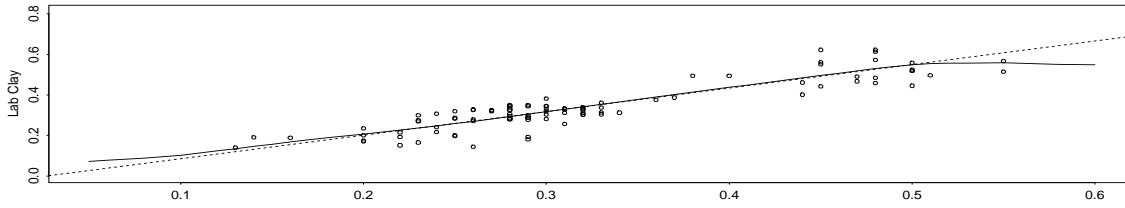
- National Cooperative Soil Survey produces soil surveys:
 - county-level soil maps and soil series descriptions
 - intensive field work; rarely updated
 - used by farmers, contractors, scientists
- Map traditionally constructed from:
 - informal model
 - subjective evaluation of topography, vegetation, etc.
 - occasional observation of soil properties at purposively-located sites
- Problems:
 - no consistent measurement protocol
 - poor distributional information

Western Iowa Pilot Project

- Soil survey update in two counties
 - statistically defensible design
 - field work commitment comparable to traditional update
 - consistent measurement protocols
- Multi-phase design: focus on two
 - surface horizon points ($N = 665$)
 - auxiliary variable x_i = field clay in surface horizon
 - lab points ($n = 97$): horizon-specific lab measurements
 - study variables y_i = components of texture (clay, sand, silt)

Lab Texture Data

- Local linear fits with $h = 0.2$ and Epanechnikov kernel:



Compositional Data

- Local linear regression:

Component	HT	REG	LLR(0.2)	LLR(opt)	h_{opt}
Clay	0.346	0.351	0.351	0.351	0.221
	(0.0129)	(0.0051)	(0.0050)	(0.0050)	
Sand	0.091	0.088	0.088	0.089	0.310
	(0.0109)	(0.0067)	(0.0064)	(0.0062)	
Silt	0.563	0.561	0.561	0.564	0.128
	(0.0075)	(0.0072)	(0.0060)	(0.0058)	
Total	1.000	1.000	1.000	1.004	

Conclusions

- New class of model-assisted survey estimators
- Good theoretical properties:
 - is ADU and design consistent
 - attains Godambe-Joshi lower bound asymptotically
 - dominates classical estimator for nonlinear populations
- Good practical properties:
 - has weighted form
 - calibrates to known control totals
 - is insensitive to bandwidth choice
- Further work
 - multiple auxiliary variables
 - multi-phase and multi-stage designs