Variance Estimation for Decision-Based Stratified Regression Estimates

Eric Slud, Census Bureau and University of Maryland

Joint work with Jun Shao and his student S. Wang

plus Yang Cheng & Carma Hogue

this talk overlaps with talk of Jun Shao, in same Session 554

United States Census Bureau

OUTLINE

O. Background – Motivation from ASPEP & previous work

I. Intro with/without Sampling, Weights, and homoscedastic-error Linear Regression in substrata

II. MSE benefits of pooling – unconditional or conditional on X

III. Bootstrap Variance for Decision-Based Estimator

IV. Other variance estimators ...

Background

Annual Survey of Public Employment and Payroll (ASPEP)

- response variables Y relating to full and part-time employment in Government units

- stratified design by State and 4 government 'Types'

- strata for **Subcounty** and **Special District** further subdivided into small and large substrata by total-payroll size variable

- sampled PPS within strata, **subsampled** within small-substrata

Model-Assisted Regression Estimation

- predictor variable X_i : same as response variable Y_i , but taken from previous Government Census (2007 Response, 2002 Predictor in dataset analyzed)

- Totals by substratum $t_X^{(sm)}, t_X^{(lg)}$ **known**

- positive size variable Z_i , PPS stratum weights w_i

- Separate estimation from 2 substrata $\hat{t}_{Y,reg}^{(sm)} + \hat{t}_{Y,reg}^{(lg)}$

- versus pooled estimator $\hat{t}_{Y,reg}^{(pool)}$ from combined substratum

Decision-Based Estimation

(Hypothesis-test-based pooling)

Substratum k $= 0, 1$ **:** size n_k, model $Y_i = a_k + b_k X_i + \epsilon_i^{(k)}$

Combined: size $n_0 + n_1$, model $Y_j = \alpha + \beta X_j + \epsilon_j$

Combination Rule: pool if $|\widehat{b}_0 - \widehat{b}_1| \le 1.96 \cdot SE$

GREG Estimators, **Variances** PPSWR or PPSWOR

Part 1, Research Issue: Is there an MSE Benefit from sub-stratum collapsing, even if a single regression model holds ?

Levels of complexity

- Sample size, large vs. small (Large in Shao's talk)
- iid sampling vs. biased sampling vs. Survey
- Linear additive-error regression vs. General model

Non-survey, Linear Regression Case

Within Substratum $k = 0, 1,$ $Y_i \sim \mathcal{N}(a_k + b_k X_i,\, \sigma_k^2)$ iid data

Sample n_k , with **known** $\mu_{k,x}$ substratum mean of X_i and **known** proportion λ_k (interpretation: $N_k/(N_0 + N_1)$)

Objective: from 2-substratum (X_i, Y_i) data, estimate Y-mean

$$E(Y) = \mu_Y = \lambda_0 \,(a_0 + b_0 \mu_{0,x}) + \lambda_1 \,(a_1 + b_1 \mu_{1,x})$$

For simplicity, assume $a_0 + b_0 c = a_1 + b_1 c,$ c *known, e.g.,* c *may be a cut-point in* X*'s used to split substrata.*

Two Statistics

Least-squares estimators $\begin{cases} \text{substratum} & \widehat{a}_k,\, \widehat{b}_k, \quad k=0,1 \\ \text{pooled} & \widehat{\alpha},\ \widehat{\beta} \end{cases}$

$$T = \sum_{k=0}^{1} \lambda_k \{\bar{Y}_k + \widehat{b}_k(\mu_{k,x} - \bar{X}_k)\}$$

$$S = \sum_{k=0}^{1} \lambda_k \{\bar{Y}_k + \widehat{\beta}(\mu_{k,x} - \bar{X}_k)\}$$

T is unbiased

Bias in S proportional to $\delta = (b_1 - b_0)/\sqrt{\sigma_0^2 + \sigma_1^2}$

Difficult to Improve Unconditional MSE by Collapsing

Simulate samples as follows (with λ_k known):

- *iid* $\mathbf{X}$ samples within substrata defined by X_i below/above cutoff c equal to quantile (usually 0.8, taken $=\lambda$), $\mu_{k,x}$ known

- Y_i's generated by equal-b_k linear regressions with normal errors in substrata, $\gamma = \sigma_0^2/(\sigma_0^2 + \sigma_1^2)$

For each pair of substratum-k samples (X_i, Y_i) of size n_k, $k = 0, 1$ find $(T - \mu_Y)^2$, $(S - \mu_Y)^2$, **Compute averages = MSE's**, also the S-bias multiple of δ.

Table of MSE's at $\delta = 0$, & Breakeven δ_*

$\lambda_0 = .8 =$ quantile for cutoff; simulations with $R = 5000$

Dist. of X	n_0	n_1	γ	relΔMSE	δ_*
$\mathcal{N}(4,1)$	100	50	.5	.0094	.0621
	50	30		.0197	.1246
	40	20		**.0261**	.1624
$\mathcal{N}(4,1)$	100	50	.25	.0129	.0567
	50	30		.0236	.1053
	40	20		**.0362**	.1502
Expon(1)	100	50	.25	.0126	.0266
	50	30		.0245	.0515
	40	20		**.0375**	.0740
Lognorm(0, 1)	100	50	.25	.0119	.0079
	50	30		.0384	.0214
	40	20		**.0444**	.0249

Idea: next consider Conditional MSE's

$MSE(T|\mathbf{X})$ versus $MSE(S|\mathbf{X})$

Notations: $\bar{X}_k\,,\ S^2_{k,x}$ stratum sample mean, var

$$\Delta = \sum_{k=0}^{1} \lambda_k\,(\mu_{k,x} - \bar{X}_k)$$

$$D = (n_0 - 1)S^2_{0,x} + (n_1 - 1)S^2_{1,x} + \frac{n_0\,n_1}{n_0 + n_1}(\bar{X}_0 - \bar{X}_1)^2$$

Conditional Bias $E(S\,|\,\mathbf{X}) - \mu_Y \ = \ \delta[\lambda_1(\bar{X}_1 - \mu_{1,x}) +$

$$+\frac{\Delta}{D}((n_1 - 1)\,S^2_{1,x} + (c - \bar{X}_1)\,(\bar{X}_0 - \bar{X}_1)\,\frac{n_0\,n_1}{n_0 + n_1}]$$

Conditional Variance Formulas

$$\mathrm{Var}(T|\mathbf{X}) = \sum_{k=0}^{1} \lambda_k^2 \, \frac{\sigma_k^2}{n_k} \, (1 + \frac{n_k \, (\mu_{k,x} - \bar{X}_k)^2}{(n_k - 1)\, S_{k,x}^2})$$

$$\mathrm{Var}(S|\mathbf{X}) = \sum_{k=0}^{1} \frac{\sigma_k^2}{n_k} (\lambda_k + (2k-1)\frac{n_0\, n_1}{n_0 + n_1} \, (\bar{X}_0 - \bar{X}_1) \, \frac{\Delta}{D})$$

$$+ \; ((n_0 - 1)\, S_{0,x}^2 \, \sigma_0^2 + (n_1 - 1)\, S_{1,x}^2 \, \sigma_1^2) \, \frac{\Delta^2}{D^2}$$

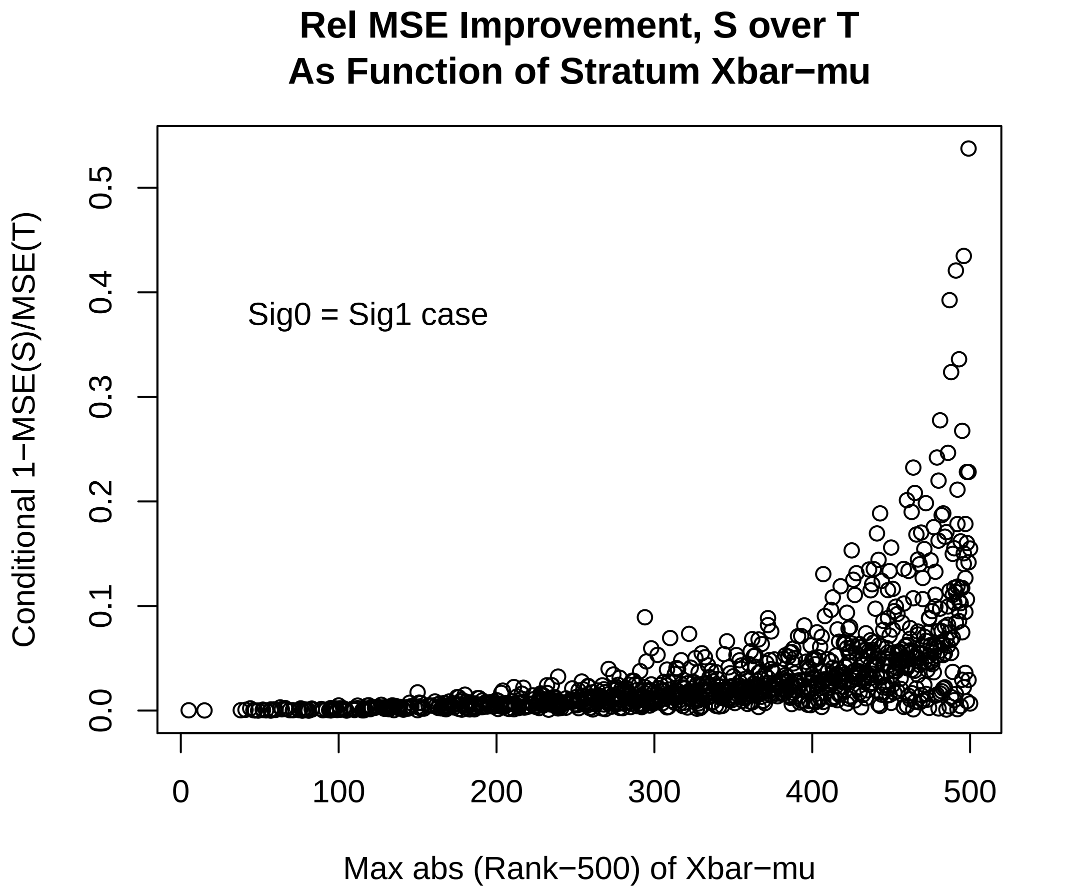
Are there aspects of X data that can tell us when conditional MSE improvements are substantial ?

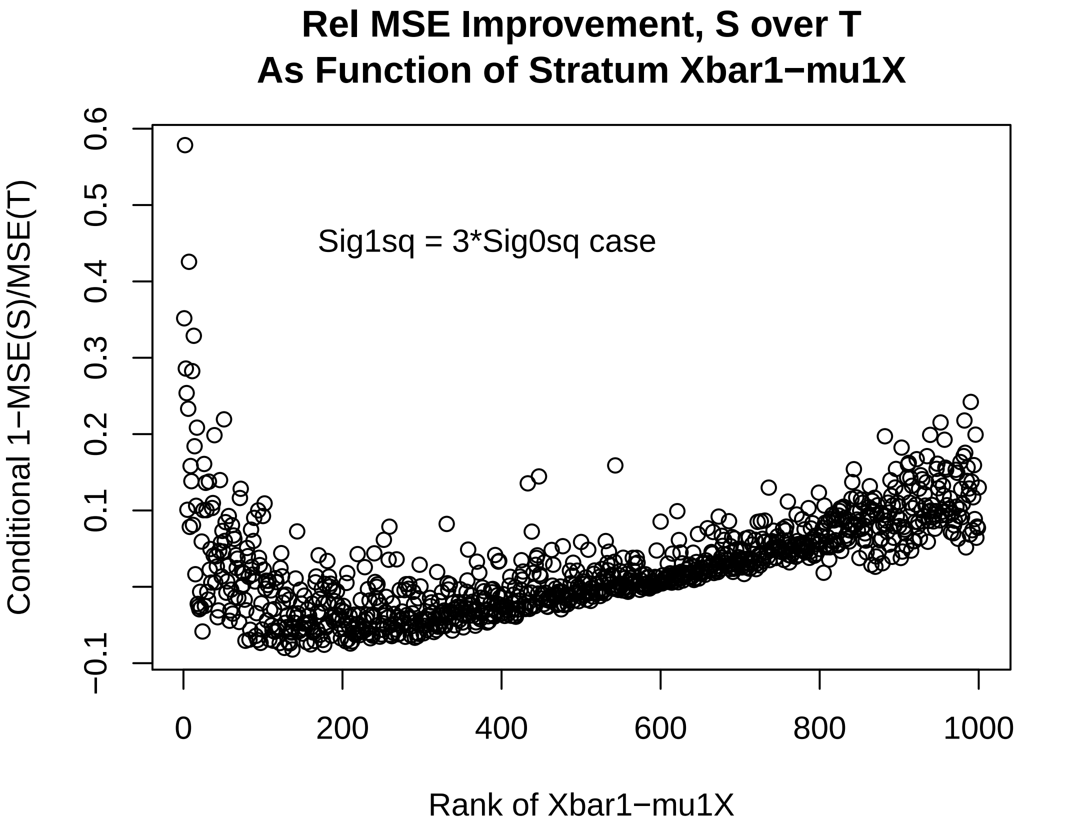
1st Graph looks at 1000 simulated LogNorm(0,1) samples

plots conditional MSE improvement $1 - MSE(S)/MSE(T)$
versus max of abs(rank-500) of $\bar{X}_k - \mu_{k,x}, \; k = 0, 1$
(normal linear regressions with $\sigma_1^2 = \sigma_0^2$)

2nd Graph: conditional MSE improvement $1 - MSE(S)/MSE(T)$
versus rank of $\bar{X}_1 - \mu_{1,x}$ ($\sigma_1^2 = 3\sigma_0^2$)

Note: occasional improvements up to 40-60% !!


Rel MSE Improvement, S over T
As Function of Stratum Xbar−mu
Sig0 = Sig1 case
Conditional 1−MSE(S)/MSE(T)
Max abs (Rank−500) of Xbar−mu
0.0
0.1
0.2
0.3
0.4
0.5
0
100
200
300
400
500


Rel MSE Improvement, S over T
As Function of Stratum Xbar1−mu1X
Sig1sq = 3*Sig0sq case
Conditional 1−MSE(S)/MSE(T)
−0.1
0.1
0.2
0.3
0.4
0.5
0.6
0
200
400
600
800
1000
Rank of Xbar1−mu1X

Tentative Conclusions, Part 1

- Meaningful MSE improvements due to hypothesis-test-based collapsing of substrata is not possible in large samples

- Unconditional MSE improvements of more than a few percent seem not to be possible in moderate and small samples

- Useful conditional MSE improvements of S over T do seem possible if substrata are combined only when $\bar{X}_k - \mu_{k,x}$ is large (usually for $k = 1$). Best form of decision-based estimator still not clear.

These statements have been confirmed also in PPS survey-sampling setting. Conditional improvements may diminish for some PPS weights.

Part 2: Can Bootstrap or some other method accurately estimate Variance of the decision-based estimates ?

Recall form of **decision-based** estimator $\hat{t}_{Y,dec}$

$$= \begin{cases} \sum_{k=0}^{1} (\hat{t}_{Y,k} + \hat{b}_k(t_{x,k} - \hat{t}_{x,k}) & \text{if } \ |\hat{b}_1 - \hat{b}_0| \le 1.96 \cdot SE \\ \hat{t}_Y + \hat{\beta}(t_X - \hat{t}_X) & \text{otherwise} \end{cases}$$

Naive: (Cheng et al. 2010) survey variance estimator of stratified survey regression estimator (combined or two-substratum) chosen by test.

Bootstrap: (i) resample equiprobably with replacement from pairs (X_i, Y_i) within each substratum;
(ii) apply complete 2-stage definition of $\hat{t}^{(b)}_{Y,dec}$ in b'th re-sample, and (iii) take resulting sample variance of $\{\hat{t}^{(b)}_{Y,dec}\}_{b=1}^{B}$.

Bootstrapping Hypothesis Tests

(Bickel & Ren 2000, Beran 1986, Shao & Tu 1995)

Bootstrap cannot estimate power of hypothesis tests !

We can see why, in a test for mean $\mu = 0$ based on *iid* data $\mathbf{Z} = \{Z_i\}_{i=1}^n$ with finite variances.

Let $\bar{Z}_n^{*(b)}$ = sample mean of b'th bootstrap sample from **Z**:

then bootstrap theory as in Shao and Tu (1995) says: with probability 1 as n gets large

$$\sqrt{n}(\bar{Z}_n^{*(b)} - \bar{Z}_n) \stackrel{\mathcal{D}}{\to} \mathcal{N}(0, \sigma_Z^2)$$

The 'natural' bootstrap estimator of power of the test rejecting for $\sqrt{n}\,|\bar{Z}_n| \ge z_{\alpha/2}\,\sigma_Z$ is: $\hat{\pi}^B = B^{-1}\sum_{b=1}^{B} I[\sqrt{n}\,|\bar{Z}^{*(b)}| \ge z_{\alpha/2}\,\sigma_Z]$

For $\mathbf{Z}$ from a law with any mean μ ,

$$P(\sqrt{n}\,|\bar{Z}_n^{*(b)}| \ge z_{\alpha/2}\,\sigma_Z\,|\,\mathbf{Z}) \to P(|W_0 + \sqrt{n}\bar{Z}| \ge z_{\alpha/2}\,\sigma_Z)\,|\,\mathbf{Z})$$

with $W_0 \sim \mathcal{N}(0, \sigma_Z^2)$ (and later indep. W_1) indep. of $\mathbf{Z}$.

So when $\mu = h/\sqrt{n}$, the expectation of $\hat{\pi}^B$ tends to

$$P(|W_0 + \sqrt{n}(\bar{Z} - \mu) + h| \ge z_{\alpha/2}\,\sigma_Z) \to P(|W_0 + W_1 + h| \ge z_{\alpha/2}\,\sigma_Z)$$

while $\hat{\pi}^B$ 'should' estimate the power $P(|W_1 + h| \ge z_{\alpha/2}\,\sigma_Z)$.

Consequences for Decision-Based Estimators

We should therefore **not** expect that straightforward bootstrap could estimate correctly the probability that $\hat{t}_{Y,dec}$ coincides with the 2-substratum stratified regression estimator.

Example 1: consider 'survey' with substrata chosen SRS ($n_0 = 50, n_1 = 30$) from populations $N_0 = 1600, N_1 = 400$, in which $X_i \sim Gamma(4, .1)$ are *iid* split at .8 quantile, and $Y_i = 20 + 1.5X_i + \epsilon_i$ in both substrata, $\epsilon_i \sim \mathcal{N}(0, 100)$.

Example 2: same except $b_1 - b_0 = 2$, $\delta = 2/\sqrt{200}$.

Simulated $R = 5000$ Monte Carlo Iterations,
with $B = 100$ bootstrap replications.

Bootstrap & Monte Carlo Simulation Results in Examples

	UnWght.1	UnWght.2	Wght.1	Wght.2
MC.pRej	.078	.149	.080	.215
Boot.pRej	.208	.268	.216	.318
True t_Y	162678	163707	160651	161400
avg ty.Dec	162699	163629	160642	161381
avg ty.2str	162692	163620	160645	161405
DecSE.emp	2440.8	2353.8	2542.0	2586.3
DecSE.Naiv	2380.1	2320.5	2437.6	2466.1
DecSE.Boot	2406.5	2344.9	2504.0	2524.1
Naiv.Boot	2420.6	2352.1	2524.1	2555.1

Conclusions

- In small samples with widely dispersed X it **can** pay to collapse substrata.

- Best to collapse when at least one $\bar{X}_k - \mu_{k,x}$ is large.

- Further research needed to explore how to exploit conditional MSE improvement selectively based on **X**.

- Bootstrap works adequately for variance in most cases although clearly **not** for power (i.e. of estimating probability of maintaining 2 substrata).

- But bootstrap seems no better than Naive method.

References

Part 1:

Barth, J., Cheng, Y., and Hogue, C. (2009), JSM 2009

Cheng, Y., Slud, E., and Hogue, C. (2010) JSM 2010

Shao, J., Slud, E., Cheng, Y., Wang, S. and Hogue, C. (2011).

Part 2:

Beran, R. (1986) Simulated power functions. Ann. Stat.

Bickel, P.J. and Ren, J. (2001), The Bootstrap in hypothesis testing. IMS Lec.Notes **36**

Shao, J. and Tu, D. (1995) **The Jackknife and Bootstrap**