A modeler's perspective on survey weights

Roderick Little

UNIVERSITY OF MICHIGAN

Outline of talk

1. Big picture: design vs. model-based inference, weighting vs. prediction
2. Comparisons of weighting, prediction and generalized regression
3. Advocate robust modeling strategies

My overarching philosophy: calibrated Bayes

- Survey inference is not fundamentally different from other problems of statistical inference
  - But it has particular features that need attention
- Statistics is basically prediction: in survey setting, predicting survey variables for non-sampled units
- Inference should be model-based, Bayesian
- Seek models that are "frequency calibrated":
  - Incorporate survey design features
  - Properties like design consistency are useful
  - "objective" priors generally appropriate
    - Little (2004, 2006); Little & Zheng (2007)

Design vs. model-based survey inference

- Design-based (Randomization) inference
  - Survey variables Y fixed, inference based on distribution of sample inclusion indicators, I
- Model-based inference: Survey variables Y also random, assigned statistical model, often with fixed parameters. Two variants:
  - Superpopulation: Frequentist inference based on repeated samples from sample and superpopulation (hybrid approach)
  - Bayes: add prior for parameters; inference based on posterior distribution of finite population quantities
- key distinction in practice is randomization or model

Design-based vs. model-based inference

Design-based estimators

- design unbiased
- potentially very inefficient
- variance estimation is cumbersome, and CI may deviate from nominal level at small sample size

Parametric model-based estimators

- subject to bias when the underlying model is misspecified
- efficient if model is correct
- variance estimation is more straightforward

Zheng and Little (2003, 2005)

Robust Bayesian predictive estimators

- robust to model misspecification
- efficient
- variance or CI is estimated from posterior distribution, and the confidence coverage is close to the nominal level

Weighting

- A pure form of design-based estimation is to weight sampled units by inverse of inclusion probabilities $w_i = 1/\pi_i$
  - Sampled unit i "represents" w_i units in the population
- More generally, a common approach is:

$$w_i = w_{is} \times w_{in}(w_{is}) \times w_{ip}(w_{is}, w_{in})$$

w_{is} = sampling weight

$w_{in}(w_{is})$ = nonresponse weight

$w_{ip}(w_{is}, w_{in})$ = post-stratification weight

Prediction

- The goal of model-based inference is to predict the non-sampled values

$$\hat{T} = \sum_{i \in s} y_i + \sum_{i \in \bar{s}} \hat{y}_i$$

$$\hat{y}_i = \text{ prediction based on model } M$$

- Prediction approach captures design information with covariates, fixed and random effects, in the prediction model
- (objective) Bayes is superior conceptual framework, but superpopulation models are also useful

Composite approaches

- Generalized Regression Estimator, e.g.

$$\overline{y}_{\mathrm{GR}} = N^{-1}\left(\sum_{i=1}^{N}\hat{y}_i + \sum_{i=1}^{n} w_i(y_i - \hat{y}_i)\right)$$

- Combines prediction and weighting
- Calibration by weighted residuals conveys robustness against model violations
- With proper attention to the model (Firth and Bennett 1998, calibration weighting is unnecessary and degrades the inference (examples below)
- Change the model, not the estimator

The common ground

- Weighters can’t ignore models
- Modelers can’t ignore weights

Weighters can’t ignore models

- Weighting units yields design-unbiased or design-consistent estimates
  - In case of nonresponse, under “quasirandomization” assumptions
- Simple, prescriptive
  - Appearance of avoiding an explicit model
- But poor precision, confidence coverage when “implicit model” is not reasonable
  - Extreme weights a problem, solutions often ad-hoc
  - Basu’s (1971) elephants

Modelers can't ignore weights

- All models are wrong, some models are useful
- Models that ignore features like survey weights are vulnerable to misspecification
  - Inferences have poor properties
  - See e.g. Kish & Frankel (1974), Hansen, Madow & Tepping (1983)
- But models <u>can</u> be successfully applied in survey setting, with attention to design features
  - Weighting, stratification, clustering

Prediction trumps weighting as a principle

- All statistics is fundamentally prediction
- It's the model that matters: best way to interpret alternative estimators is to consider the model assumptions, and the implied predictions
- Weighted estimates imply predictions, but the weights implied by prediction models have no inherent interpretation, except in special cases
- Consider for example regression prediction

Ex 1: regression prediction

Regression estimate of mean of Y based on auxiliary variable Z

$$\bar{y}_{\text{reg}} = \bar{y} + \hat{\beta}_{y \cdot x}(\bar{z} - \bar{Z}) = \sum_{i=1}^{n} w_i y_i,$$

$$\text{where } w_i = \frac{1}{n} + \frac{(\bar{z} - \bar{Z})(z_i - \bar{z})}{\sum_{\ell=1}^{n} (z_\ell - \bar{z})^2}$$

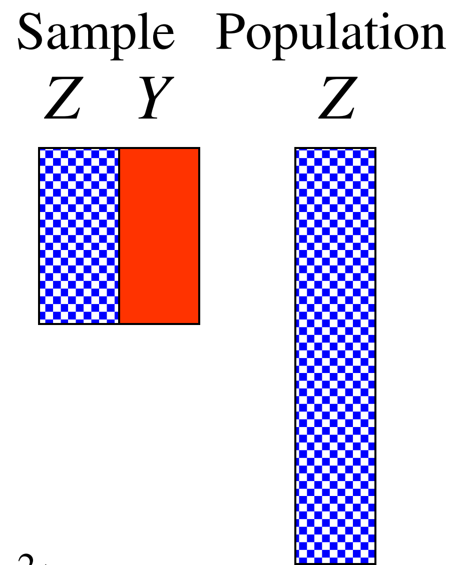
Performance of estimator depends on validity of model assumptions

Is the relationship between Y and Z linear? etc.

Weights are simply a by-product of the estimator:

Form of the weights provides no insight on whether or not this is a good model

Ex 2. One categorical post-stratifier Z



$$\bar{y}_{\text{reg}} = \bar{y}_{\text{wt}} = \sum_{j=1}^{J} P_j \bar{y}_j = \sum_{j=1}^{J} w_j n_j \bar{y}_j \Big/ \sum_{j=1}^{J} w_j n_j$$

In post-stratum j:

$P_j =$ population proportion

$n_j =$ sample count, $\bar{y}_j =$ sample mean of Y

$\bar{y}_{\text{reg}} =$ prediction estimate for $y_{ji} \sim \text{Nor}(\mu_j, \sigma_j^2)$

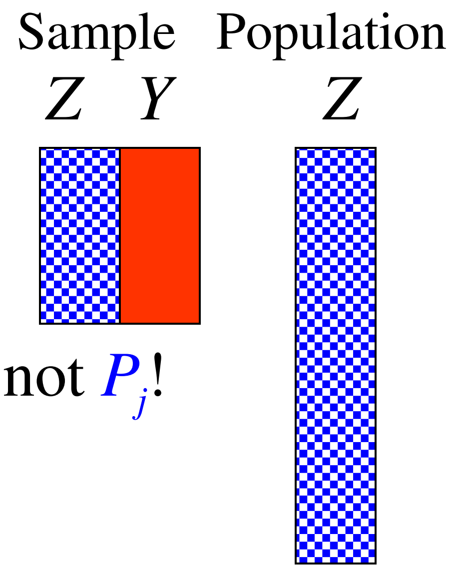
$w_j = nP_j / n_j =$ implied weight in poststratum j

This weight has an interpretation – adjusting the sample population to the known population proportion in post-stratum j

One categorical post-stratifier Z

Weight focus Prediction focus

$$\bar{y}_{\text{reg}} = \bar{y}_{\text{wt}} = \sum_{j=1}^{J} P_j \bar{y}_j = \sum_{j=1}^{J} w_j n_j \bar{y}_j \Big/ \sum_{j=1}^{J} w_j n_j$$



Estimator breaks down in small samples:

Design modifies weight w_j -- but problem is with $\bar{y}_j$, not P_j!

Model replaces $\bar{y}_j$ by prediction $\hat{\mu}_j$ from model

Changing $\bar{y}_j$ requires a model --

E.g. $\mu_j \sim \text{Nor}(\mu, \tau^2)$ shrinks weight towards 1.

Implied weights from this random effects model depend on values of Y

Model assumptions are more informative than the form of the weights

Model vs GR estimator

Random Effects Model:

$$\hat{\mu}_{\mathrm{RE}} = \sum_{j=1}^{J} \tilde{w}_j n_j \bar{y}_j \,/\, \sum_{j=1}^{J} \tilde{w}_j n_j ,\ \tilde{w}_j :$$

$\tilde{w}_j$ shrinks w_j towards 1, by amount dependent on closeness of poststratum means

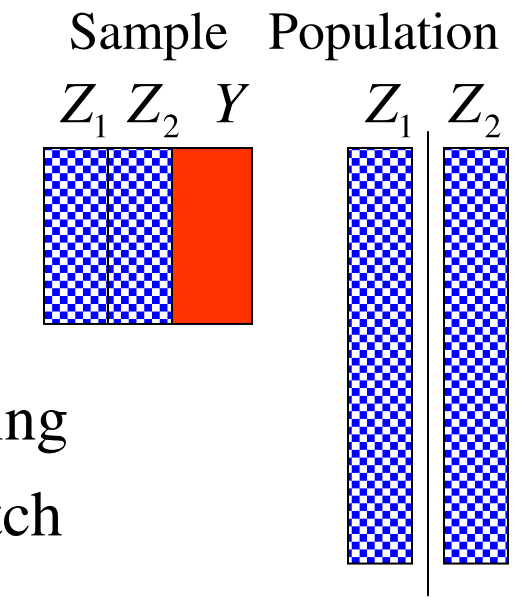
Generalized Regression Estimator does not help here:

$$\hat{\mu}_{\mathrm{GR}} = \bar{y} + \sum_{j=1}^{J} w_j n_j (\bar{y}_j - \bar{y}) \,/\, \sum_{j=1}^{J} w_j n_j = \bar{y}_{\mathrm{wt}} :$$

Reduces to design-weighted estimator, no gain from model

Ex 3. One stratifier Z_1, one post-stratifier Z_2

<u>Design-based approaches</u>



(A) Standard weighting is $w_i = w_{is} \times w_{ip}(w_{is})$

Notes: (1) Z_1 proportions are not matched!

(2) why not $w_i^* = w_{ip} \times w_{is}(w_{ip})$?

(B) Deville and Sarndal (1987) modifies sampling weights $\{w_{is}\}$ to adjusted weights $\{w_i\}$ that match poststratum margin, but are close to $\{w_{is}\}$ with respect to a distance measure $d(w_{is}, w_i)$.

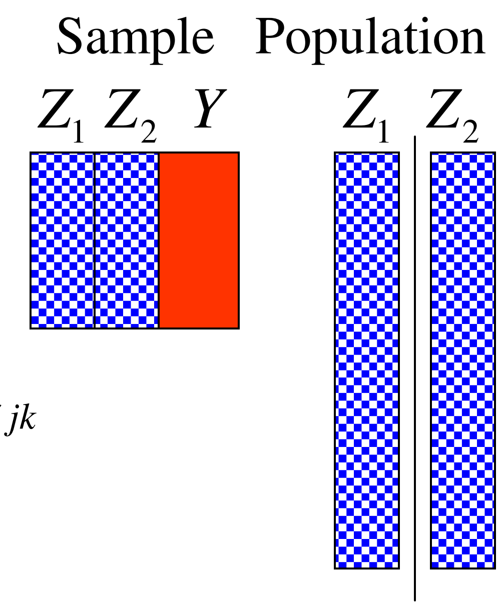
Questions:

What is the principle for choosing the distance measure?

Should the $\{w_i\}$ necessarily be close to $\{w_{is}\}$?

Ex 3. One stratifier Z_1, one post-stratifier Z_2

Model-based prediction approach



Saturated model: $\{n_{jk}\} \sim \text{MNOM}(n, \pi_{jk})$;

$$y_{jki} \sim \text{Nor}(\mu_{jk}, \sigma^2_{jk})$$

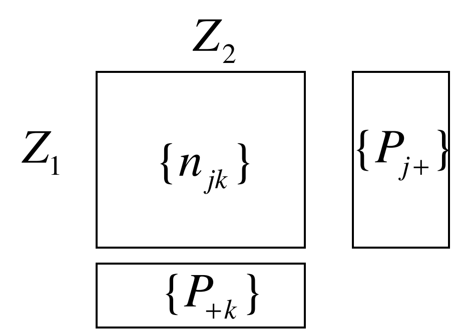
$$\bar{y}_{\text{mod}} = \sum_{j=1}^{J}\sum_{k=1}^{K} \hat{P}_{jk}\bar{y}_{jk} = \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk} n_{jk} \bar{y}_{jk} \Big/ \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk} n_{jk}$$

$n_{jk} =$ sample count, $\bar{y}_{jk} =$ sample mean of Y

$\hat{P}_{jk} =$ proportion from raking (IPF) of $\{n_{jk}\}$

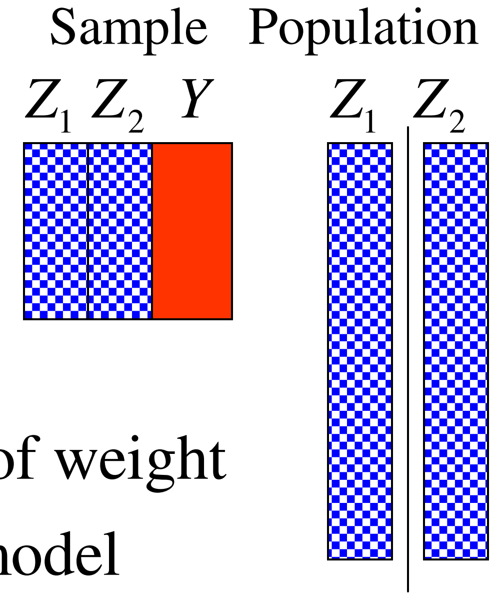
to known margins $\{P_{j+}\}, \{P_{+k}\}$

$w_{jk} = n\hat{P}_{jk} / n_{jk} =$ model weight



Ex 3. One stratifier Z_1, one post-stratifier Z_2

<u>Model-based approach</u>



$$\bar{y}_{\text{mod}} = \sum_{j=1}^{J}\sum_{k=1}^{K} \hat{P}_{jk}\bar{y}_{jk} = \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk} n_{jk} \bar{y}_{jk} / \sum_{j=1}^{J}\sum_{k=1}^{K} w_{jk} n_{jk}$$

What to do when n_{jk} is small?

Design: arbitrary collapsing, ad-hoc modification of weight

Model: Replace $\bar{y}_{jk}$ by prediction from modified model

e.g. $y_{jki} \sim \text{Nor}(\mu + \alpha_j + \beta_k + \gamma_{jk}, \sigma^2_{jk})$,

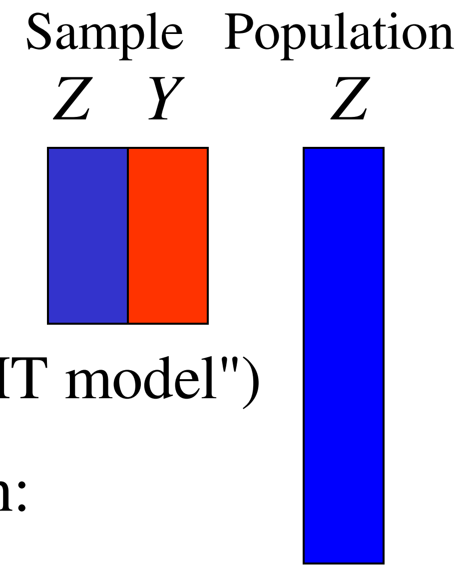
$\sum_{j=1}^{J}\alpha_j = \sum_{k=1}^{K}\beta_k = 0,\ \gamma_{jk} \sim \text{Nor}(0, \tau^2)$ (Gelman 2007)

Setting $\tau^2 = 0$ yields additive model,

otherwise shrinks towards additive model

Ex 4. One continuous (post)stratifier Z

Consider PPS sampling, Z = measure of size



<u>Design</u>: HT or Generalized Regression

$$\bar{y}_{\text{wt}} = \frac{1}{N}\left(\sum_{i=1}^{n} y_i / \pi_i\right); \pi_i = \text{ selection prob (HT)}$$

$$\bar{y}_{\text{wt}} \approx \text{ prediction estimate for } y_i \sim \text{Nor}(\beta\pi_i, \sigma^2\pi_i^2) \text{ ("HT model")}$$

This motivates following robust modeling approach:

$$\bar{y}_{\text{mod}} = \frac{1}{N}\left(\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{y}_i\right), \ \hat{y}_i \text{ predictions from:}$$

$$y_i \sim \text{Nor}(S(\pi_i), \sigma^2\pi_i^k); S(\pi_i) = \text{penalized spline of } Y \text{ on } Z$$

Zheng and Little (2003, 2005) show reduced RMSE, better confidence coverage than HT, GR estimators by simulation:

Probit p-spline regression model

- Chen, Elliott and Little (2009) extend p-spline model to a probit model for a binary outcome (Ruppert, Wand, and Carroll 2003):

$$\Phi^{-1}\left(P\left(y_i = 1\right)\right) = \beta_0 + \sum_{k=1}^{p} \beta_k \pi_i^k + \sum_{l=1}^{m} b_l \left(\pi_i - k_l\right)_+^p$$

$$b_l \sim N\left(0, \tau^2\right) \quad l = 1,\ldots,m \quad i = 1,\ldots,n$$

  - the constants $k_1 < \ldots < k_m$ are m selected fixed knots.
  - $(u)_+^p = \{u \times I(u \geq 0)\}^p$ for any real number u.
  - Gibbs' sampling used to generate posterior predictive distribution of nonsampled values
  - 95% credible interval: split the tail area of posterior distribution equally between the upper and lower 2.5% endpoints.

Simulation study (1)

- **Unequal probability sampling design:**
  - PPS sampling: units are selected with probability proportional to a given size variable related to the survey variable under study.
- **Population and sample:**
  - N = 2,000 with sampling rates of 5% and 10% (n = 100 or 200).
  - The size variable *X* takes the values 71, 72, ..., 2070. The inclusion probabilities π are proportional to *X*.
- **Simulations:** 1000 replicates
- **Compare:**
  - Empirical bias, width of posterior probability/CI
  - Empirical root mean squared error (RMSE)
  - Noncoverage rate of 95% CI

Simulation study (1): artificial populations

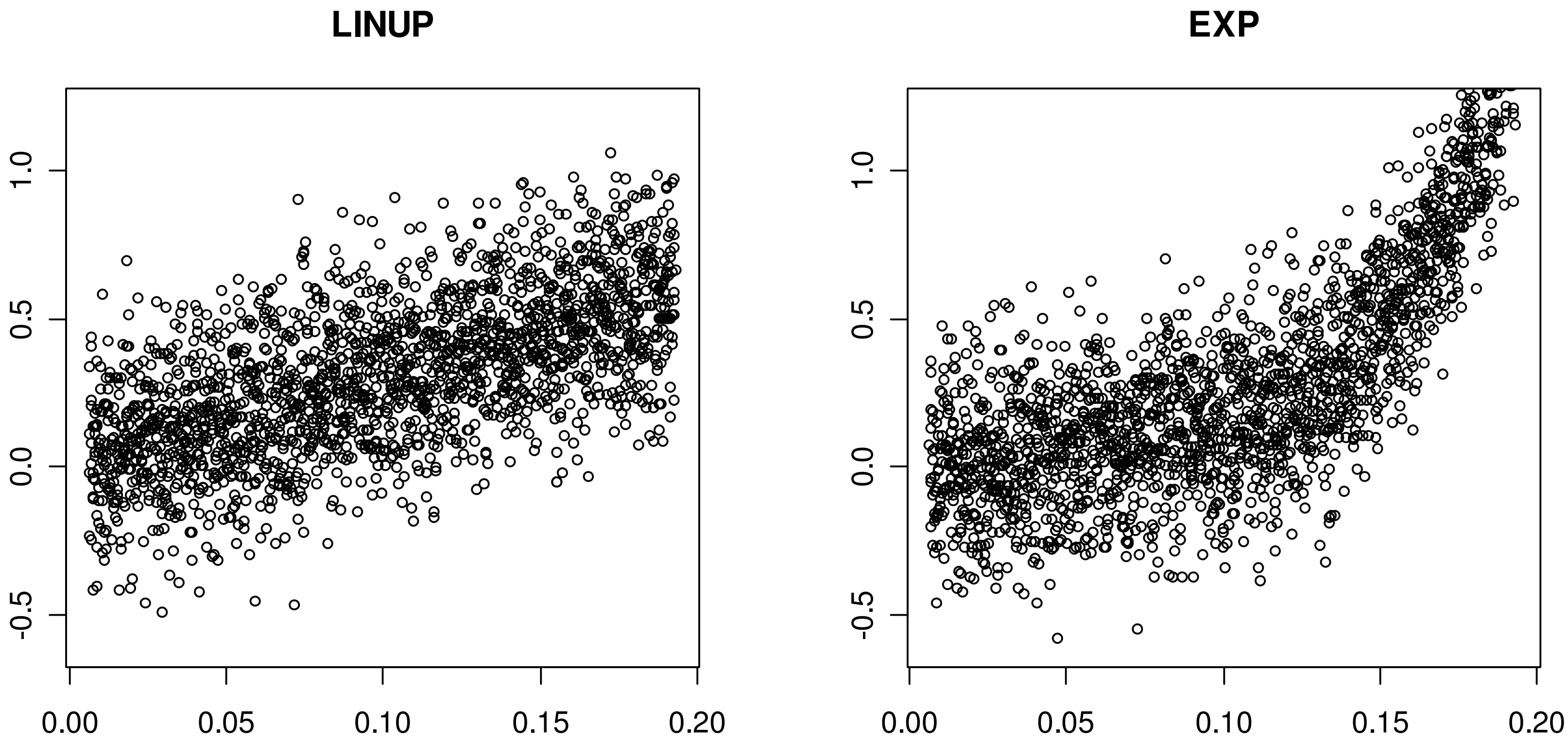


Figure 1 Two simulated populations, X-axis: inclusion probability; Y-axis: Y*

Simulation study (1): RMSE (low=good)

Table 1 Empirical RMSE × 1,000 of six estimators (Minimum RMSE is in bold print)

Population	True prop.	HK	LR	PR	PR_GR	BPSP	BPSP_GR
LINUP N=2,000 n=100	0.10	55	57	**47**	52	**47**	52
	0.50	66	52	**48**	51	49	51
	0.90	27	24	**24**	24	**24**	24
EXP N=2,000 n=100	0.10	**52**	60	55	**52**	**52**	53
	0.50	67	57	**44**	54	48	53
	0.90	25	**13**	**13**	**13**	**13**	**13**

★ BPSP method yields estimates with small RMSE

Simulation study (1): CI noncoverages (nominal = 5)

Table 2 Noncoverage rate of 95% CI × 100 of six estimators (noncoverage rate most close to 5 is in bold print)

Population	True prop.	HK	LR		PR	PR_GR		BPSP	BPSP_GR	
			V1	V2		V1	V2		V1	V2
LINUP N=2,000 n=100	0.10	16.7	24.7	18.5	**8.3**	21.8	16.0	8.9	18.9	14.2
	0.50	7.3	6.9	10.8	5.7	7.5	9.6	**5.4**	7.9	9.0
	0.90	7.9	8.3	11.1	7.0	8.8	9.2	**6.8**	8.8	9.3
EXP N=2,000 n=100	0.10	14.8	24.7	17.7	10.9	19.2	14.9	**9.7**	18.5	14.3
	0.50	8.9	8.7	12.8	12.5	9.1	10.4	**8.3**	10.5	10.0
	0.90	**6.7**	12.2	12.2	9.7	12.3	9.1	9.7	11.2	9.0

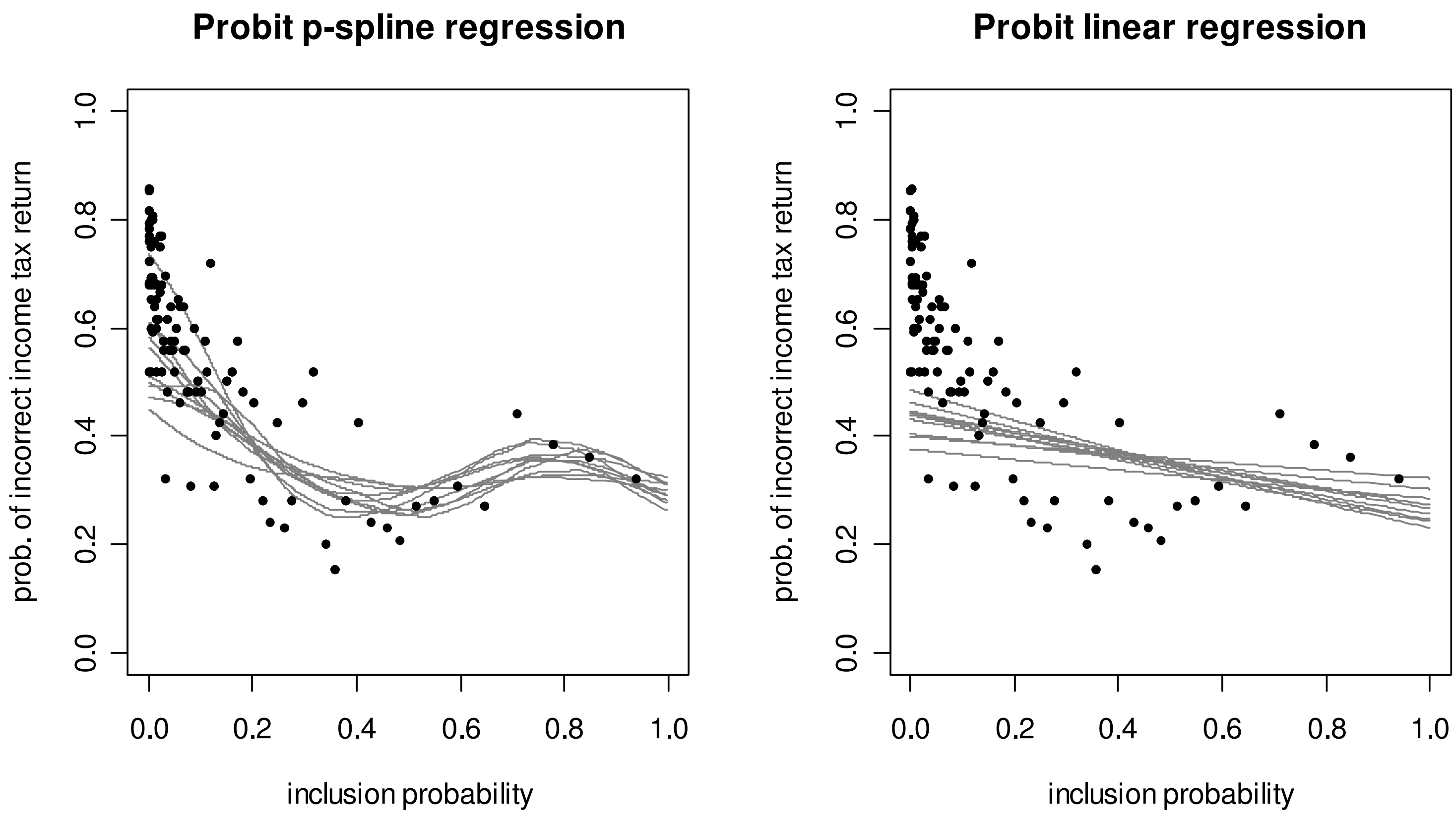
* V1: variance estimator using linearization; V2: jackknife variance estimator

BPSP method has confidence coverage closer to nominal level, especially when true prop. = 0.10

Simulation study (2)

- Tax auditing data (Compumine 2007)
  - 3,119 income tax returns
  - *Y:* whether the income tax return is incorrect (p=0.517)
  - *X*: the amount of the realized profit
  - PPS sampling using *X* as the size variable
  - n = 300 or 600
  - 1,000 replicates of simulation

Simulation study (2): Tax auditing data



Simulation study (2): results

Table 3 Comparison of various estimators for empirical bias, root mean squared error, and average width and noncoverage rate of 95% CI, in the tax return example

Methods	bias*100		RMSE*100		average width*100		noncoverage*100	
	300	600	300	600	300	600	300	600
HK	-2.4	-1.8	12.4	10.2	36	29	14.1	10.2
LR	6.7	5.5	11.9	9.2	27	21	43.5	45.6
PR	-11.6	-10.1	12.4	10.6	18	14	69.8	83.4
PR_GR	-1.2	-0.3	11.5	8.8	33	26	16.1	11.4
BPSP	-6.8	-2.7	9.3	5.2	27	19	14.2	5.0
BPSP_GR	-0.7	0.2	12.0	10.1	34	26	15.9	12.8

* The variance of GR estimator is estimated using linearization

- BPSP estimator performs well; PR estimator is biased and has poor confidence coverage because of model misspecification

Why does model do better?

- Assumes smooth relationship – HT weights can "bounce around"
- Predictions use sizes of the non-sampled cases
  - HT estimator does not use these
  - Often not provided to users (although they could be)
- Little & Zheng (2007) also show gains for model when sizes of non-sampled units are not known
  - Predicted using a Bayesian Bootstrap (BB) model
  - BB is a form of stochastic weighting

Summary

- Compared design-based and model-based approaches to survey weights
- Model approach is attractive because of flexibility, inferential clarity
- Advocate survey inference under “weak models”

References

Chen, Q., Elliott, M. and Little, R. (2009). Under review, *Survey Meth*.

Deville, J-C and Sarndal, C-E. JASA 87, 376-382

Firth, D. & Bennett, K.E. (1998). *JRSS* B 60, 3-21.

Gelman, A. (2007) *Stat. Science* (with discussion)

Hansen, MH, Madow, WG & Tepping, BJ (1983) *JASA* 78, 776-793.

Holt, D., & Smith, T.M.F. (1979). *JRSS*A, 142, 33-46.

Horvitz, D.G., & Thompson, D.J. (1952). *JASA*, 47, 663-685.

Kish, L., & Frankel, M. R. (1974). *JRSS* B, 36, 1–37.

Little, R.J.A. (2004). *JASA,* 99, 546-556.

______ (2006). *Am. Statist.*, 60, 3, 213-223.

______ & Zheng, H. (2007). To appear in *Bayesian Statistics* 8, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith & M. West (Eds.)

Rizzo, L. (1992). *JASA* 87, 1166-1173.

Rosenbaum, P. R. & Rubin, D. B. (1983). *Biometrika*, 70, 41-55

Rubin, D.B. (1985). pp. 463-472 in *Bayesian Statistic*s, eds. I.M. Bernado, M.H. DeGroot, D.V. Lindley, & A.F.M. Smith, Amsterdam: Elsevier,.

Särndal, C.-E., Swensson, B. & Wretman, J.H. (1992), *Model Assisted Survey Sampling*, Springer: New York.

Zheng, H. & Little, R.J. (2003). *JOS* 19, 2, 99-117.

______ (2005) *JOS* 21, 1-20.