Should we use the survey weights to weight?

Roderick Little

University of Michigan
Outline of talk
1. Big picture: design vs. model-based inference, weighting vs. prediction
2. Comparisons of weighting and prediction
3. Weighting and prediction for nonresponse
4. Robust modeling strategies
5. Variance estimation and inference
Outline of talk

1. Big picture: design vs. model-based inference, weighting vs. prediction
2. Comparisons of weighting and prediction
3. Weighting and prediction for nonresponse
4. Robust modeling strategies
5. Variance estimation and inference
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**
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
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

• Compare weighting and prediction approaches, and argue for model-based prediction
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
Ex 1. Basu’s inefficient elephants

\( (y_1, ..., y_{50}) = \) weights of \( N = 50 \) elephants

Objective: \( T = y_1 + y_2 + ... + y_{50} \). Only one elephant can be weighed!

- Circus trainer wants to choose “average” elephant (Sambo)
- Circus statistician requires “scientific” prob. sampling:
  Select Sambo with probability 99/100
  One of other elephants with probability 1/4900
  Sambo gets selected! Trainer: \( \hat{t} = y_{(Sambo)} \times 50 \)

Statistician requires unbiased Horvitz-Thompson (1952) estimator:

\[
\hat{T}_{HT} = \begin{cases} 
  y_{(Sambo)} / 0.99 (!!); \\
  4900 y_{(i)}, \text{if Sambo not chosen (!!!)}
\end{cases}
\]

HT estimator is unbiased on average but always crazy!
Circus statistician loses job and becomes an academic
What went wrong?

• HT estimator optimal under an implicit model that $y_i / \pi_i$ have the same distribution
• That is clearly a silly model given this design …
• Which is why the estimator is silly
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 can be successfully applied in survey setting, with attention to design features
  – Weighting, stratification, clustering
Outline of talk

1. Big picture: design vs. model-based inference, weighting vs. prediction
2. Comparisons of weighting and prediction
3. Weighting and prediction for nonresponse
4. Robust modeling strategies
5. Variance estimation and inference
Ex 2. One categorical post-stratifier \( Z \)

\[
\bar{y}_{\text{mod}} = \bar{y}_{\text{wt}} = \frac{\sum_{j=1}^{J} P_j \bar{y}_j}{\sum_{j=1}^{J} w_j n_j} = \frac{\sum_{j=1}^{J} w_j n_j \bar{y}_j}{\sum_{j=1}^{J} w_j n_j}
\]

In post-stratum \( j \):

\( P_j \) = population proportion

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

\( w_j = nP_j / n_j \)

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

\( = \) design-weighted estimate with weight \( w_j \)

Hence two approaches intersect ...
One categorical post-stratifier $Z$

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

Sample $Z$ Population $Z$

1. Approaches differ in small samples:
   - **Model** replaces $\bar{y}_j$ by prediction $\hat{\mu}_j$ from model
   - E.g. $\mu_j \sim \text{Nor}(\mu, \tau^2)$ shrinks weight towards 1.
   - **Design** modifies weight $w_j$ -- but problem is with $\bar{y}_j$, not $P_j$!
   - Changing $\bar{y}_j$ requires a model --
   - Modifications of weights often have an implicit model
   - Preferable to be explicit!
One categorical post-stratifier $Z$

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

2. Variance estimation:

Model automatically fixes the counts \(\{n_j\}\);

Bayes with objective prior yields t-type corrections

Design approach is not clear on whether to condition on \(\{n_j\}\)

\(\{n_j\}\) are not fixed in repeated sampling

But if allowed to vary, the sampling variance is not defined!

(Holt and Smith 1971, Little 1993)
Ex 3. One stratifier $Z_1$, one post-stratifier $Z_2$

**Design-based approaches**

(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 (1992) 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 $\{w_{is}\}$?

Survey weights

Sample | Population
---|---
$Z_1$ | $Z_2$ | $Y$ | $Z_1$ | $Z_2$
Ex 3. One stratifier $Z_1$, one post-stratifier $Z_2$

Model-based approach

Saturated model: \( \{n_{jk}\} \sim \text{MNOM}(n, \pi_{jk}) \);
\[ y_{jki} \sim \text{Nor}(\mu_{jk}, \sigma_{jk}^2) \]

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

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

\( \hat{P}_{jk} = \) proportion from raking (IPF) of \( \{n_{jk}\} \)
to known margins \( \{P_{j+}\}, \{P_{+k}\} \)

\( w_{jk} = n_{jk} \hat{P}_{jk} / n_{jk} = \) model weight
Ex 3. One stratifier $Z_1$, one post-stratifier $Z_2$

Model-based approach

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

What to do when $n_{jk}$ is small?

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

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

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

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

Design: arbitrary collapsing, ad-hoc modification of weight

Survey weights
Ex 4. One continuous (post)stratifier $Z$

Consider PPS sampling, $Z =$ measure of size

**Design:** HT or Generalized Regression

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

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

This motivates following robust modeling approach:

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

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

(Zheng and Little 2003, 2005)
Simulation: PPS sampling in 6 populations

Survey weights
## Estimated RMSE of four estimators for N=1000, n=100

<table>
<thead>
<tr>
<th>Population</th>
<th>model</th>
<th>wt</th>
<th>gr</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>Normal</td>
<td>20</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Lognormal</td>
<td>32</td>
<td>44</td>
</tr>
<tr>
<td>LINUP</td>
<td>Normal</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Lognormal</td>
<td>25</td>
<td>30</td>
</tr>
<tr>
<td>LINDOWN</td>
<td>Normal</td>
<td>30</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>Lognormal</td>
<td>24</td>
<td>65</td>
</tr>
<tr>
<td>SINE</td>
<td>Normal</td>
<td>35</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td>Lognormal</td>
<td>53</td>
<td>130</td>
</tr>
<tr>
<td>EXP</td>
<td>Normal</td>
<td>26</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Lognormal</td>
<td>40</td>
<td>41</td>
</tr>
</tbody>
</table>
## 95% CI coverages: HT

<table>
<thead>
<tr>
<th>Population</th>
<th>V1</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>90.2</td>
<td>91.4</td>
<td>90.0</td>
<td>90.4</td>
</tr>
<tr>
<td>LINUP</td>
<td>94.0</td>
<td>95.0</td>
<td>95.0</td>
<td>95.0</td>
</tr>
<tr>
<td>LINDOWN</td>
<td>89.0</td>
<td>89.8</td>
<td>90.0</td>
<td>90.6</td>
</tr>
<tr>
<td>SINE</td>
<td>93.2</td>
<td>93.4</td>
<td>93.0</td>
<td>93.0</td>
</tr>
<tr>
<td>EXP</td>
<td>93.6</td>
<td>94.6</td>
<td>95.0</td>
<td>95.0</td>
</tr>
<tr>
<td>ESS</td>
<td>95.0</td>
<td>95.6</td>
<td>95.4</td>
<td>95.2</td>
</tr>
</tbody>
</table>

V1  Yates-Grundy, Hartley-Rao for joint inclusion probs.
V3  Treating sample as if it were drawn with replacement
V4  Pairing consecutive strata
V5  Estimation using consecutive differences
# 95% CI coverages: B-spline

<table>
<thead>
<tr>
<th>Population</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>95.4</td>
<td>95.8</td>
<td>95.8</td>
</tr>
<tr>
<td>LINUP</td>
<td>94.8</td>
<td>97.0</td>
<td>94.6</td>
</tr>
<tr>
<td>LINDOWN</td>
<td>94.2</td>
<td>94.2</td>
<td>94.6</td>
</tr>
<tr>
<td>SINE</td>
<td>88.0</td>
<td>92.6</td>
<td>97.4</td>
</tr>
<tr>
<td>EXP</td>
<td>94.4</td>
<td>95.2</td>
<td>95.6</td>
</tr>
<tr>
<td>ESS</td>
<td>97.4</td>
<td>95.4</td>
<td>95.8</td>
</tr>
</tbody>
</table>

V1  Model-based (information matrix)  
V2  Jackknife  
V3  BRR

Survey weights
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
Outline of talk

1. Big picture: design vs. model-based inference, weighting vs. prediction
2. Comparisons of weighting and prediction
3. Weighting and prediction for nonresponse
4. Robust modeling strategies
5. Variance estimation and inference
Ex 5. Unit nonresponse

• Predict nonrespondents by regression on design variables $Z$ and any observed survey variables $X$

• For bias reduction, predictors should be related to propensity to respond $R$ and outcome $Y$

• In choosing from a set of predictors, good prediction of $Y$ is more important than good prediction of $R$
Survey weights

Impact of weighting for nonresponse

\[ \text{corr}^2(X,Y) \]

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>---</td>
<td>var ↓</td>
</tr>
<tr>
<td>High</td>
<td>var ↑</td>
<td>var ↓</td>
</tr>
</tbody>
</table>

Too often adjustments do this?

- Standard “rule of thumb” \( \text{Var}(\bar{Y}_w) = \text{Var}(\bar{Y}_u)(1 + \text{cv}(w)) \)
  fails to reflect that nonresponse weighting can **reduce** variance
- Little & Vartivarian (2005) propose refinements
Weighting squared?

- Nonresponse weights are often computed using units weighted by their sampling weights $\{w_{1i}\}$
  
  $$w_{2j}^{-1} = \left( \sum_{r_i=1, x_i=j} w_{1i} \right) / \left( \sum_{r_i=1, x_i=j} w_{1i} + \sum_{r_i=0, x_i=j} w_{1i} \right)$$

- Gives unbiased estimate of response rate in each adjustment cell defined by $X$
- Not correct from a prediction perspective
- For nonresponse bias, need to condition on $X$ and $Z$, not just $X$
- Does not correct for bias when $Z$ is associated with $R$ and $Y$
- Need to condition on design variables involved in sampling weight (as in predictive inference)
Simulation Study

- Simulations to provide insight into the variance and bias of estimators of weighted and unweighted rates and alternative estimators, under a variety of population structures and nonresponse mechanisms. (Little & Vartivarian 2003)

- Categorical outcome, to avoid distributional assumptions such as normality.

- 25 populations to cover all combinations of models for $Y$ and $R$ given $Z$ and $X$
### RMSE’s of three methods

<table>
<thead>
<tr>
<th>Y model</th>
<th>XZ</th>
<th>XZ</th>
<th>XZ</th>
<th>XZ</th>
<th>add</th>
<th>add</th>
<th>add</th>
<th>add</th>
<th>Z</th>
<th>Z</th>
<th>Z</th>
<th>Z</th>
<th>Ave*</th>
</tr>
</thead>
<tbody>
<tr>
<td>R model</td>
<td>XZ</td>
<td>add</td>
<td>X</td>
<td>Z</td>
<td>Ø</td>
<td>XZ</td>
<td>add</td>
<td>X</td>
<td>Z</td>
<td>Ø</td>
<td>Ave*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pred(xz)</td>
<td>40</td>
<td>41</td>
<td>42</td>
<td>36</td>
<td>36</td>
<td>38</td>
<td>39</td>
<td>39</td>
<td>37</td>
<td>33</td>
<td>36</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>urr(x)</td>
<td>49</td>
<td>53</td>
<td>42</td>
<td>51</td>
<td>37</td>
<td>54</td>
<td>71</td>
<td>40</td>
<td>59</td>
<td>34</td>
<td>58</td>
<td>76</td>
<td>39</td>
</tr>
<tr>
<td>wrH(x)</td>
<td>51</td>
<td>58</td>
<td>42</td>
<td>53</td>
<td>37</td>
<td>56</td>
<td>77</td>
<td>41</td>
<td>61</td>
<td>35</td>
<td>56</td>
<td>71</td>
<td>39</td>
</tr>
</tbody>
</table>

(*Models for $Y$ that exclude $Z$ are omitted to save space – methods are all similar for these cases)

- $urr(x)$ is biased when both $Y$ and $R$ depend on $Z$.
- $wrr(x)$ does not generally correct the bias in these situations: similar to $urr[x]$ overall

Prediction based on model for $X$ and $Z$ is best

Survey weights 31
Item nonresponse

- Item nonresponse generally has complex “swiss-cheese” pattern
- Weighting methods are possible when the data have a monotone pattern, but are very difficult to develop for a general pattern
- Model-based multiple imputation methods are available for this situation (Little & Rubin 2002)
  - By conditioning fully on all observed data, these methods weaken MAR assumption
Outline of talk

1. Big picture: design vs. model-based inference, weighting vs. prediction.
2. Comparisons of weighting and prediction
3. Weighting and prediction for nonresponse
4. Robust modeling strategies
5. Variance estimation and inference
Making predictions more robust

- Model predictions of missing values are potentially sensitive to model misspecification, particularly if data are not MCAR.
Relaxing Linearity: one $X$

- A simple way is to categorize $X_1$ and predict within classes -- link with weighting methods
- For continuous $X_1$ and sufficient sample size, a spline provides one useful alternative (cf. Breidt & Opsomer 2000). We use a P-Spline approach:

$$(Y_1 | X_1, \phi) \sim \text{Nor} \left( s_1(X_1, \phi), \sigma^2 \right)$$

$$s_1(X_1, \phi) = \phi_0 + \sum_{j=1}^{q} \phi_j X_1^j + \sum_{k=1}^{K} \phi_{q+k} \left( X_1 - \tau_k \right)_+^q,$$

$$(x)_+^q = x^q I(x \geq 0),$$

$\tau_1 < \cdots < \tau_K$ are selected fixed knots

$\phi_{q+1}, \ldots, \phi_{q+K}$ are random effects, shrink to zero

Survey weights
More than one covariate

• When we model the relationship with many covariates by smoothing, we have to deal with the “curse of dimensionality”.
  – One approach is to “calibrate” the model by adding weighted residuals (e.g. Scharfstein & Izzarry 2004, Bang & Robins 2005).
  – Strongly related to generalized regression approach in surveys (Särndal, Swensson & Wretman 1992).
Penalized Spline of Propensity Prediction (PSPP)

• Focus on a particular function of the covariates most sensitive to model misspecification, the response propensity score.

• Important to get relationship between $Y$ and response propensity correct, since misspecification of this leads to bias (Rubin 1985, Rizzo 1992)

• Other $X$’s balanced over respondents and nonrespondents, conditional on propensity scores (Rosenbaum & Rubin 1983); so misspecification of regression of these is less important (loss of precision, not bias).
The PSPP method

Define: \( Y^* = \text{logit} \left( \Pr(R = 1 | X_1, \ldots, X_p) \right) \) (Need to estimate)

\[
(Y | Y^*, X_1, \ldots, X_p; \beta) \sim \text{Nor}(s(Y^*) + g(Y^*, X_2, \ldots, X_p; \beta), \sigma^2)
\]

- Nonparametric part
- Need to be correctly specified
- We choose penalized spline

- Parametric part
- Misspecification does not lead to bias
- Increase precision
- \( X_1 \) excluded to prevent multicollinearity

Survey weights
Double Robustness Property

- The PSPP method yields a consistent estimator for the marginal mean of $Y$, if:
  
  (a) the mean of $Y$ given $X_1, \ldots, X_p$ is correctly specified,

  OR

  (b1) the propensity is correctly specified, and
  (b2) relationship of outcome with propensity is correctly specified

Note: in (b), parametric part of model does not have to be correctly specified!

- PSPP can be extended to handle general patterns of missing data

- Applies to other problems of selection, e.g. sampling (Little & An 2004)
Role of Models in Classical Approach

• Models are often used to motivate the choice of estimator. For example:

  – Regression model → regression estimator
  – Ratio model → ratio estimator
  – Generalized Regression estimation: model estimates adjusted to protect against misspecification, e.g. HT estimation applied to residuals from the regression estimator (e.g. Särndal, Swensson & Wretman 1992).

• Estimates of standard error are then based on the randomization distribution

• This approach is design-based, model-assisted
Comments

• Calibration approach yields double robustness
• However, relatively easy to achieve double robustness in the direct prediction approach, using methods like PSPP (see Firth & Bennett 1998)
• Calibration estimates can be questionable from a modeling viewpoint
• If model is robust, calibration is unnecessary and adds noise
  – Recent simulations by Guanyu Zhang support this
Outline of talk

1. Big picture: design vs. model-based inference, weighting vs. prediction.
2. Comparisons of weighting and prediction
3. Weighting and prediction for nonresponse
4. Model-assisted estimation, double robustness
5. Variance estimation and inference
Standard errors, inference

• Should be more emphasis on confidence coverage, less on estimating the standard error

• Model-based inferences
  – Need to model variance structure carefully
  – Bayes: good for small samples

• Sample reuse methods (bootstrap, jackknife, BRR)
  – More acceptable to practitioners
  – Large sample robustness (compare sandwich estimation)
  – Inferentially not quite as pure, but practically useful
Summary

• Compared design-based and model-based approaches to survey weights
• Design-based: “VW beetle” (slow, reliable)
• Model-based: “T-bird” (more racy, needs tuning)
• Personal view: model approach is attractive because of flexibility, inferential clarity
• Advocate survey inference under “weak models”
Acknowledgments

• JPSM for invitation to speak
• Don Rubin for inspiration and guidance
• Trivellore Raghunathan for being a great colleague
• Current and past graduate students for all their ideas and work
  – Di An, Hyonggin An, Michael Elliott, Laura Lazzaroni, Hui Zheng, Sonya Vartivarian, Mei-Miau Wu, Ying Yuan, Guangyu Zhang
References

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