Notes and Comments

Joe Sedransk
JPSM, CWRU and JSSAM

Rust: Design-based Inference

- Weighting is Important
  - Data files for large scale surveys contain weights
  - Secondary data users
- Weighting is Difficult
  - Inferential problems with survey data due to complex designs, nonresponse (unit, item), coverage issues, *measurement errors*
  - Ex. (Rust #9) For sample NR adjustment difficulty in finding variables that are consistently measured for R and NR *and* correlated with both outcome and response

Design-Based Methods

- *Weighting is a Mess* (first sentence of Gelman 2007)
  - Multiple uses to mitigate effect of biases, reduce variance: implies need for compromise
  - Ex (#11). Why focus on association between covariates and responses rather than outcomes? Many outcomes – but only one response variable
  - Ad hoc methods to *adjust* survey weights for NR or coverage errors, to reduce variances through use of auxiliary data or by restricting range of the weights, i.e., weight trimming.

Weight Trimming

- Design based methods differ in how cutoff boundary to identify outlying weights chosen
  - Ad hoc methods with cutoff such as median (w) + 6 IQR (w)
  - Methods using empirical MSE of estimator of interest
  - Methods that assume a specified parametric (skewed) distribution for weights

References to Models

- To increase precision of estimation (#2)
- Weight adjustments in repeated surveys using past survey data (#23)
- Response propensity models (#29)
- Modeling relationship of outcome and auxiliary variables and then using results to determine weighting approach (#29)

“Methods for Adjusting Survey Weights”

- Design-based, Model-based (predominant) and Model-assisted methods
- Goal: Make inferences robust to anomalous values of weights or y’s or both
- Approaches
  - Smooth or trim the weights
  - Smooth or trim the y’s
  - Use nonparametric estimators minimally affected by outlying weights, y’s or combinations of the two

Weight Trimming

- Often done without considering analysis variables
- Can be inefficient
  - If outlying y or wy causes estimator to have very large variance, weight trimming alone may not correct problem
  - Values of w's or y's innocuous for full pop ests may be influential for domain estimates

Weight Smoothing

$$\hat{T}_B = E_M(\hat{T}_{HT} \mid \mathbf{I}, \mathbf{Y}) = \sum_{i \varepsilon S} E_M(w_i \mid \mathbf{I}, \mathbf{Y})\, y_i = \sum_{i \varepsilon S} \tilde{w}_i\, y_i$$

Unified Approach

Model Based

Limitations

Establishment Surveys

Household Surveys

Experience With Modeling: Short-Cuts

Valliant, Dever and Kreuter "Practical Tools for Designing and Weighting Survey Samples" 2013

Valliant Presentation at US Census Bureau

- Distributions used in survey inference
- Use of models in sample design
- Use of models in constructing estimators
- Complications

Probability Distributions

- Superpopulation model
- Random selection model
- Coverage model
- Response model
- Imputation model
- Measurement error model
- Prior/Hyperprior
- Posterior

Inferential Methods

- Design based – randomization distribution
- Model based – superpopulation model
- Model assisted – models used to construct estimators; randomization distribution for inference

- Design based inference alone is not possible because of coverage errors, unit NR, item NR

Models

- Use of models inevitable and unavoidable.
- Fixation on weights rather than estimators leads us away from thinking in terms of models.
- Making models explicit clarifies procedures and makes them more understandable.
- Examining designs and estimators using models makes clear when they do and don’t work well.

Model Building

- Difficulty in finding adequate models varies depending on population and variable
  - Establishment populations
  - Household populations
- Continuous vs. categorical variables
- Main effects vs. Main + Interactions
  - Especially important in imputation models

Models and Weights

- Model based predictive inference inevitably leads to “survey weights.”
- Think of ideal model fit to survey data and consequent survey weights.
- Compare ideal survey weights with conventional survey weights.

Small Area Inference for Binary Variables in the NHIS

- Malec, Sedransk, Moriarity and LeClere. 1997 JASA
- Model development and weights

- Y_{ijk}: Cluster i, Class k, Individual j.
  $i=1,...,L;\ k=1,...,B;\ j=1,...,N_{ik}$
- Pr{$Y_{ijk} = 1 \mid p_{ik}$) = p_{ik}
- $X_k^t = (X_{k1}, \ldots, X_{kM})$, same for each individual in k
- logit (p_{ik}) = $X_k^t \beta_i$
- $\beta_i \sim N(G_i \eta, \Gamma)$
- $p(\eta, \Gamma)$ = constant

Log-Odds of Proportion

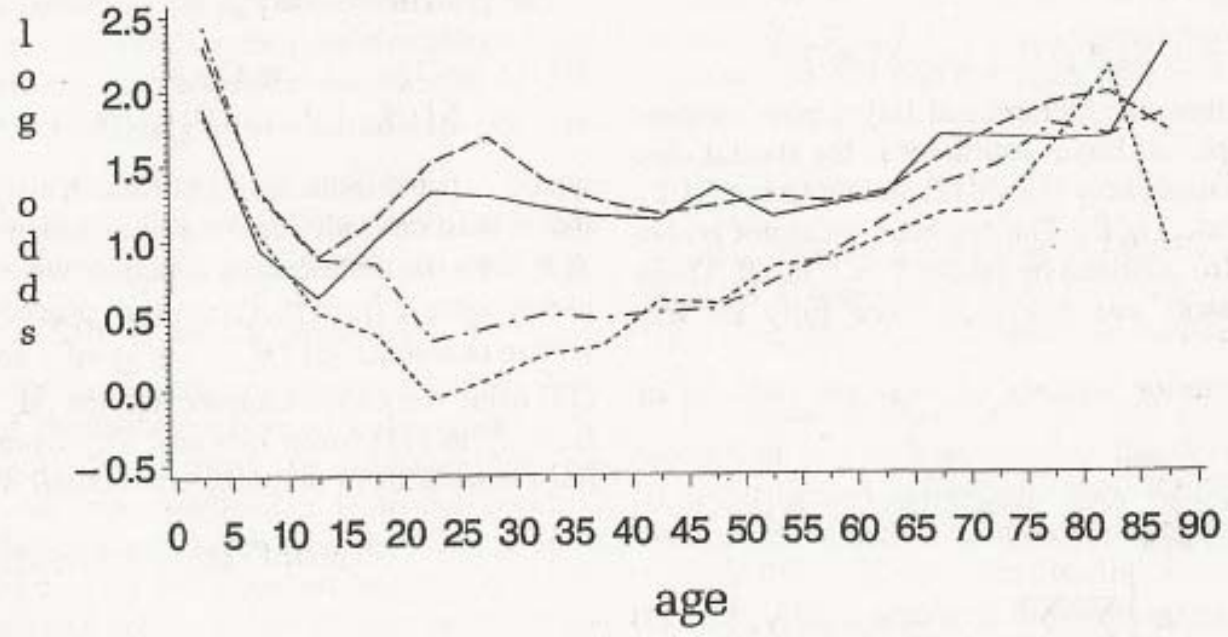


Figure 1. Relationship of Log-Odds of the Presence/Absence of at Least One Doctor Visit Within the Past Year and Age, for Each Race by Sex Class. ———, nonwhite female; - - -, nonwhite male; – – –, white female; – – – – –, white male.

Model

- logit (p_{ik}) = $\alpha + \beta_{i1}X_{0k} + \beta_{i2}X_{15,k} + \beta_{i3}X_{25,k} + \beta_{i4}X_{55,k}$
  $+ \beta_{i5}Y_kX_{15,k} + \beta_{i6}Y_kX_{25,k} + \beta_{i7}Z_k$
- Y_k and Z_k are (0,1) variables.
- Y_k = 1 if class k ~ males
- Z_k = 1 if class k ~ whites
- X_{ak} = max { 0, age k -a};
  age k ~ midpoint of ages of individuals in class k

Inference

Inference about finite population proportion

- P = $\{\sum_{i\epsilon I} \sum_{k\epsilon K} \sum_{j=1}^{N_{ik}} Y_{ikj}\}$ / $\{\sum_{i\epsilon I} \sum_{k\epsilon K} N_{ik}\}$
- Numerator of P
- $\sum_{i\epsilon I} \sum_{k\epsilon K} \sum_{j\epsilon s_{ik}} y_{ikj}$ + $\sum_{i\epsilon I} \sum_{k\epsilon K} \sum_{j \notin s_{ik}} Y_{ikj}$
- E(Numerator | y_s) = $\sum_{i\epsilon I} \sum_{k\epsilon K} \sum_{j\epsilon s_{ik}} y_{ikj}$ + $\sum_{i\epsilon I} \sum_{k\epsilon K} (N_{ik} - n_{ik}) E(p_{ik} \mid y_s)$

Model Fit

Start by ignoring variation among clusters.

- Fit logit (p_k) = X_k^tβ where p_k = Pr(Y_{ikj} = 1 | p_k).

- Plot estimate of logit (p_k) against age for(gender, race) classes
  - Probability higher for whites than non-whites for given (gender, age)
  - Patterns similar for both races for given gender
  - Males: probability decreases until age 22.5, then increases steadily
  - Females: probability decreases steadily until age 12.5, increases up to age 27.5, roughly constant until 62.5, then increases steadily.

- Fit piecewise linear spline models, i.e., linear in age
  - Race, Gender, Gender x Race
  - All interactions between these categorical variables and linear age splines

County Level Covariates

- Combine individual and county level models

  logit $(p_k) = G_i X_k^t \eta$

  Consider only seven individual level variables

  Force intercept and seven individual level vbles into model.

  Use stepwise regression to add (county level) vbles

Quality of Inferences

- Cross-Validation : Doctor Visit
- Population Based Assessment : Health-related Partial Work Limitation

Individual: Each respondent randomly allocated to one of five mutually exclusive, exhaustive groups. No controls.

County: Each county in sample randomly allocated …. No controls.

Cross-Validation Method

S_{-h} : Set of sample elements *without* those in h-th group; h = 1,…,5.

S_{ch} : Set of sample elements in h-th group in county c.

$\bar{y}_{ch} = \sum_{i \epsilon S_{ch}} y_i \; / \; |S_{ch}|$

Estimator: $E\,(\,\bar{y}_{ch} \mid S_{-h}\,)$

Compare

$D_{ch}^2 = \{E\,(\bar{y}_{ch} \mid S_{-h}) - \bar{y}_{ch}\}^2$

$V_{ch} = E\,(D_{ch}^2 \mid S_{-h})$

$C^2 = \frac{\sum_c \sum_{h=1}^5 V_{ch}}{\sum_c \sum_{h=1}^5 D_{ch}^2}$

Cross-Validation

Age	Race	Sex	Indiv	County
All	Both	Both	.99	1.05
All	Both	Female	.99	1.03
All	Both	Male	.96	1.00
All	White	Both	.99	1.03
All	Nonwhite	Both	.94	0.98
0-19	Both	Both	.96	1.07
20-64	Both	Both	.99	0.96
65+	Both	Both	.96	0.94

Sampling Weights

- Partial residual plots showed no evidence that the weights should be added as a covariate
- We have post-stratified by age, race, sex within each county and used pop weights
- All county-level variables used as stratification variables in NHIS were considered for inclusion in model
- County pop size didn’t enter model, even though it is a component of sampling weight since sampling is roughly proportional to pop size

Valliant: Conclusions

- Design unbiasedness or consistency alone does not mean good inference
- Explicit use of models for design and estimation
- Fieldwork adjustments like responsive design create design weights that may have extreme variation
- THINK ABOUT MODELING – not weighting
  - This is hard: weights often have to be done before y's are available for analysis
  - Modeling can interfere with time schedule
  - Same model doesn't work for all y's

Analytical Uses of Survey Data

- Background
- Analyst's interest: Relation of Y to X
  - Pr(Doctor visit) to age, race, sex
  - Y = Income; X = Extent of participation in voc ed program
  - Y = ln (Gross rent); Covariates: ln (HH income), ln(HH size), type of household

    Income elasticity of household expenditures
- Must take into account how survey data obtained
  - Suppose noninformative sampling and no nonresponse

    Model strata and cluster effects

    More? Weights?

Literature: Informative Sampling

- Krieger, Pfeffermann. 1992. SM
- Chambers, Dorfman, Wang. 1998. JRSS-B
- Pfeffermann, Krieger, Rinott. 1998. Sinica
- Pfeffermann, Sverchkov. 1999. Sankhya B
- Malec, Davis, Cao. 1999. Stat Med
- More recent literature, almost all frequentist, in "Inference Under Informative Sampling" by Pfeffermann, Sverchkov in *Sample Surveys: Inference and Analysis,* 29B.

Selection Bias: Ma, Nandram, Sedransk

- Structure same as CDW
  - Bayesian using full likelihood
  - Inference for any finite population quantity
  - Credible intervals
- Likelihood

Define E{ $\pi_i \mid y_i , x_i\} = \pi (y_i , x_i)$

g $(y_s, I \mid X) = \prod_{i \epsilon s}[\pi(y_i , x_i) f_p(y_i \mid x_i)/\pi_{0i}] \{ \pi_{i \notin s} (1 - \pi_{0i}) \prod_{i \epsilon s} \pi_{0i}\}$

π_{0i} = Pr(i ∈ sample | x_i) X = {x_i : i ∈ U}

"More limited pdf" h($y_s \mid X_s, s, \theta$)

Notation for Selection

$I_j = 1$, **if unit j is in sample;** $I_j = 0$, **otherwise**

$\mathbf{Pr(I)} = \prod_{i \in s} \Pi_i \prod_{i \notin s} (1 - \Pi_i)$

$\Pi_i = \frac{n\nu_i}{\sum_{j=1}^{N} \nu_j}$. $\quad \sum_{j=1}^{N} \nu_j = t$, $\quad$ **assumed known**

Model

$\nu_i = \beta_0 + \beta_1 Y_i + e_i, \qquad i = 1, ..., N$

$ln(Y_i) \sim N(\mu, \sigma^2)$

Transformation

$z_1 = \nu_1 - \bar{\nu}, ..., z_{N-1} = \nu_{N-1} - \bar{\nu}, z_N = \bar{\nu} = t/N$

$-(t/N) \leq z_i \leq (t/n) - (t/N), \quad i = 1, ..., N-1$

$[(t/N) - (t/n)] \leq \sum_{j=1}^{N-1} z_j \leq (t/N)$

Simulation Studies

- Compare NIG, IG, HT
- Inference for finite population total: Point estimator and nominal 95% intervals
- Methodology

  Fix superpopulation parameters

  Draw finite pop (N = 100) from joint distn

  Sample of size 10 using systematic pps

  Repeat 200 times

NIG: Use methodology described

HPD and equal tailed intervals

1000 Gibbs samples for approx posterior

200 samples to implement SIR

IG: Standard methodology, i.e., no selection bias

1000 Gibbs samples

HT: Usual point estimator and 95% interval

Also considered larger n, larger number of finite popIns, larger number of Gibbs/SIR samples

Results

- Plots of sample mean vs. non-sample mean
- Relative bias (average over poplns)
- Interval coverage
- Interval width (average over poplns)

Bias (Relative Mean)

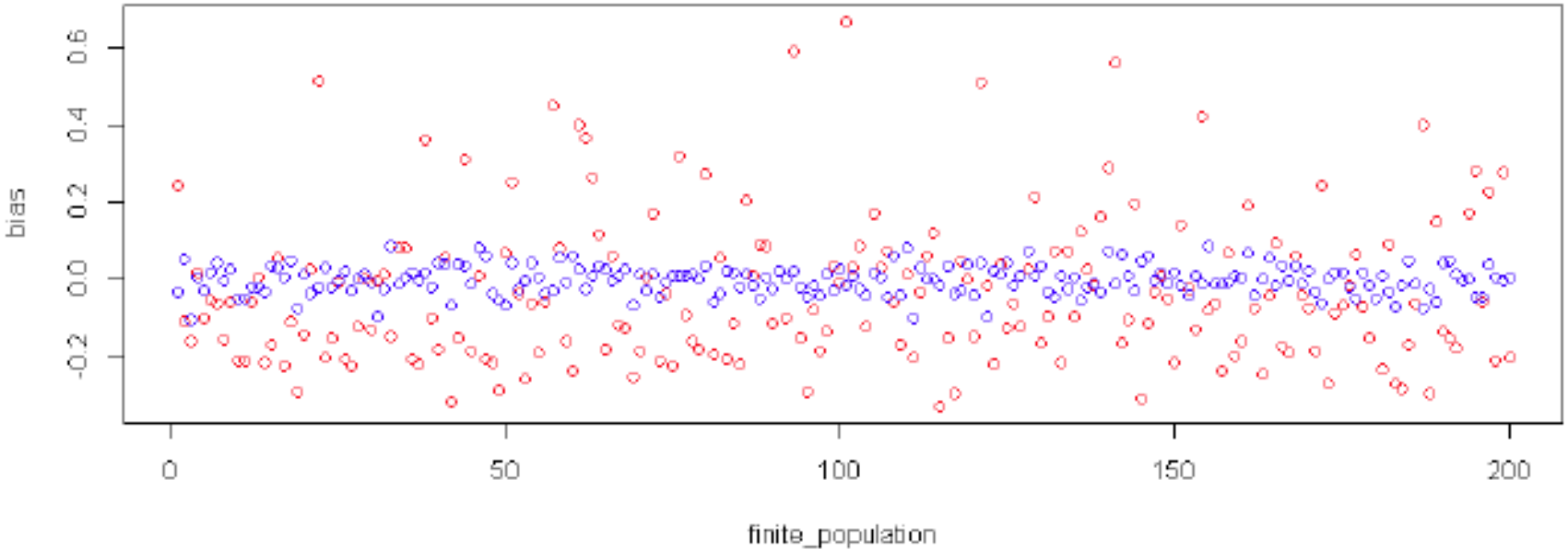
Specifications			
μ	1	1	1
σ	0.10	0.25	0.50
Corr(Y,V)	0.26	0.58	0.85
E(Y)			
IG	0.01	0.08	0.38
NIG	0.00	0.01	-0.01

Actual Coverage of Nominal 95% Interval

μ	1	1	1
σ	0.10	0.25	0.50
Corr(Y,V)	0.26	0.58	0.85
E(Y)			
IG (width)	0.37	1.09	4.38
IG (coverage)	0.92	0.94	0.97
NIG (width)	0.33	0.77	1.13
NIG (coverage)	0.89	0.94	0.94

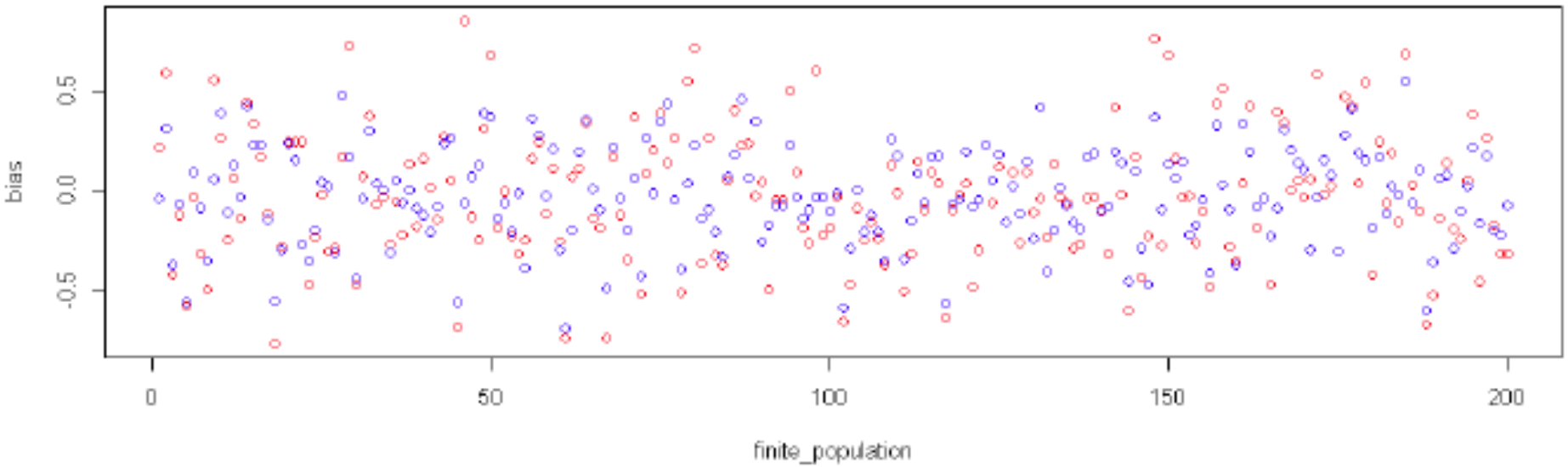
Relative Bias HT & NIG

$\beta_o = 0, \sigma = 0.04, \mu = 1$



Relative Bias HT & NIG

$\beta_o = 50, \sigma_e = 1.0$



Conclusions

- General: Models; Importance of good inference
- Specific:

  NIG corrects for selection bias

  Overall: NIG preferable to HT in important situations

  Conditional: HT has (much) greater variation in bias

  HT's best performance when Y, V proportional