Reducing Nonresponse Bias through Responsive Design and External Benchmarks

Julia Lee

University of Michigan

July 17, 2012

Thesis committee: S. Heeringa, R. Little, T. Raghunathan, R. Valliant

Goals of the Project

1. To improve respondent representativeness
2. To assess the nature of nonresponse
3. To adjust for nonresponse

Outline

- Introduction
- The proposed method
- Simulation results
- Next steps

Current Practice

Reduce nonresponse bias at the analysis stage:

- Weighting class methods
- Propensity score methods
- Calibration
- (Imputation)

Challenges:

- Need nonrespondent information
- Assume ignorable nonresponse pattern
- Extreme and highly variable weights occur

Alternatives

Reduce nonresponse bias at the design and data collection stages:

- Actively control for nonresponse bias at design stage by adaptively improving respondent representativeness.
- Effectively use frame data, contextual data, paradata, and benchmark information to obviate the need for nonrespondent information.

Responsive Design Procedure

Objectives:

- Obviate the need for nonrespondent information
- Obtain more representative respondent pool

Terminology:

- Benchmark survey: capture desired target population, such as American Community Survey
- Current Survey: survey that you are conducting

Responsive Design Procedure

Setting: Surveys with multi-phase data collection

The procedures:

1. Complete first phase of data collection.
2. Combine with benchmark information.
3. Augment with frame data, contextual data, and paradata.
4. Model the origin of each data point (1=benchmark, 0= current survey) in terms of covariates.
5. Compute ratio of propensity score density (R_{ps}) between benchmark and current survey.
6. Sample next phase subjects using R_{ps}.
7. Iterate steps 2 through 6 until acceptable representativeness or budget reached.

The problem

How do we know propensity scores of next phase subjects before they respond?

Data structure

		Y1	Y2	Y3	Y4	X1	X2	X3	Z1	Z2
Bench	**1**	√	√	√	√	√	√	√	√	√
Bench	**1**	√	√	√	√	√	√	√	√	√
Bench	**1**	√	√	√	√	√	√	√	√	√
…	**1**	√	√	√	√	√	√	√	√	√
S1	**0**	√	√	√	√	√	√	√	√	√
S1	**0**	√	√	√	√	√	√	√	√	√
…	**0**	√	√	√	√	√	√	√	√	√
S2	**0**					√	√	√	√	√
S2	**0**		Missing			√	√	√	√	√
S2	**0**					√	√	√	√	√
S2	**0**		data			√	√	√	√	√
…	**0**					√	√	√	√	√

Notation:
Ys are survey variables
Xs are common covariates across benchmark survey and the sample survey.
Zs are auxiliary data or contextual data from frame, registry, or interview observations, etc.

The key step 1: Imputation

Estimate propensity score of next samples using imputed covariates

		Y1	Y2	Y3	Y4	X1	X2	X3	Z1	Z2
Bench	**1**	√	√	√	√	√	√	√	√	√
Bench	**1**	√	√	√	√	√	√	√	√	√
Bench	**1**	√	√	√	√	√	√	√	√	√
…	**1**	√	√	√	√	√	√	√	√	√
S1	**0**	√	√	√	√	√	√	√	√	√
S1	**0**	√	√	√	√	√	√	√	√	√
…	**0**	√	√	√	√	√	√	√	√	√
S2	**0**	▲	▲	▲	▲	√	√	√	√	√
S2	**0**	▲	▲	▲	▲	√	√	√	√	√
S2	**0**	▲	▲	▲	▲	√	√	√	√	√
S2	**0**	▲	▲	▲	▲	√	√	√	√	√
…	**0**	▲	▲	▲	▲	√	√	√	√	√

Imputation

Notation:
Ys are survey variables
Xs are common covariates across benchmark survey and the sample survey.
Zs are auxiliary data or contextual data from frame, registry, or interview observations, etc.

The key step 2: R_{ps}

Define an acceptance/rejection process on the original sampling frame, to reduce or eliminate bias relative to the benchmark survey. Must satisfy:

$$\pi P(Z|accept) + (1 - \pi)P(Z) = P_B(Z)$$

where π is the fraction of the combined data that are newly drawn.

What we want is $P(accept|Z)$. Choose $P(Z)$ to be propensity score density and use Bayes Theorem to obtain

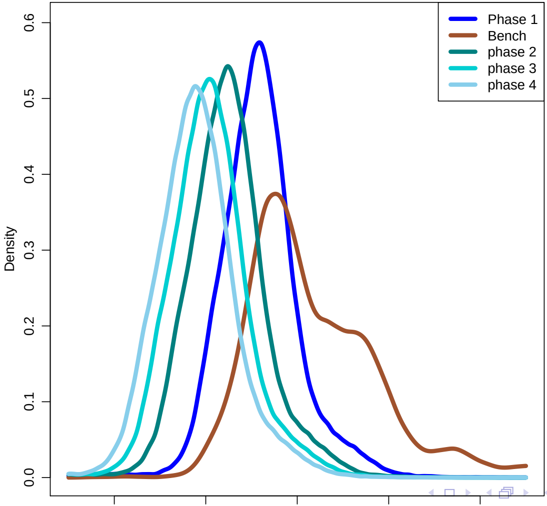
$$P(accept|Z) \propto \frac{P_B(Z)}{P(Z)}$$

NHIS vs BRFSS: Covariates in the propensity score model

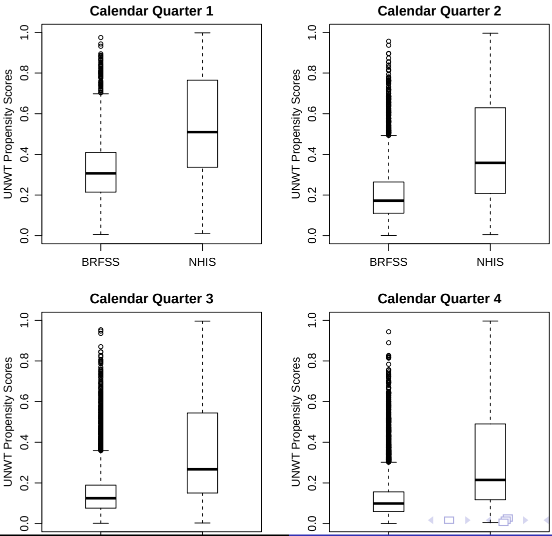
The usual suspects:

- Geographic region
- Demographic: gender, age, race, marital status,
- Socio-economic status: education, income categories, work status

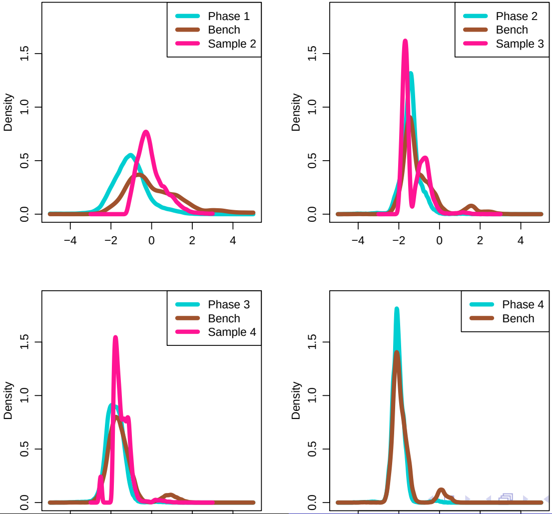
NHIS vs BRFSS: Observed Data



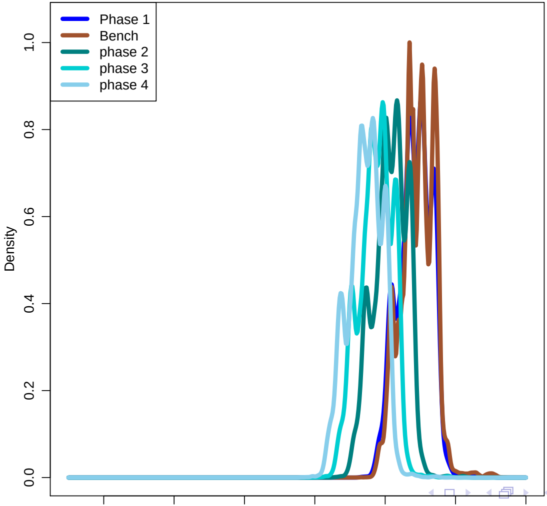
NHIS vs BRFSS: Observed Data



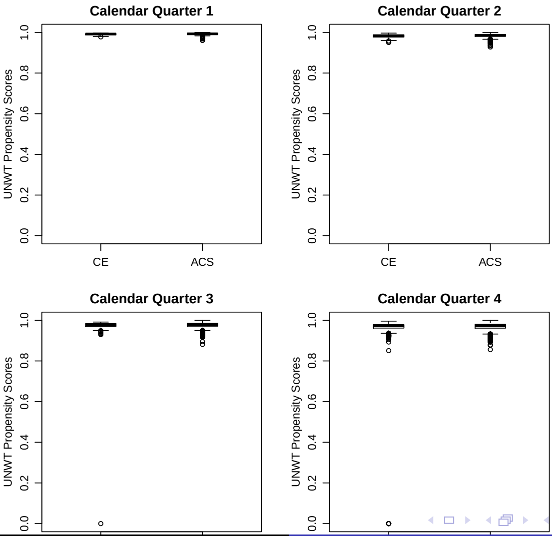
NHIS vs BRFSS: Responsive Design



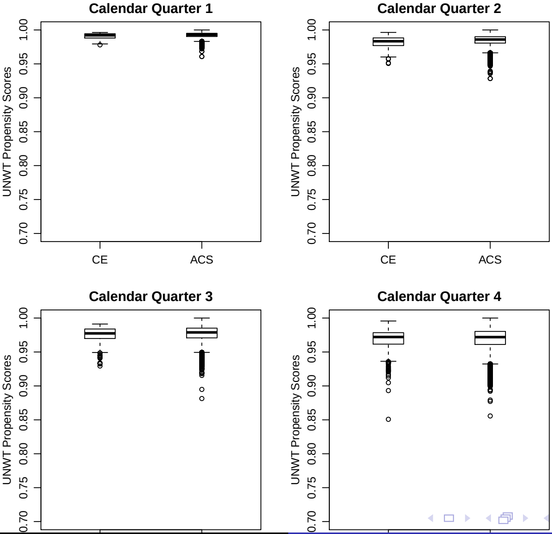
ACS vs CE: Observed Data



ACS vs CE: Observed Data



ACS vs CE: Observed Data



Model fit and similarity measures

- Model fit diagnostics
- Distance measure on densities
  - Hellinger distance to quantify the similarity between two probability distributions

$$H^2 = \frac{1}{2} \int \left(\sqrt{dP} - \sqrt{dQ} \right)^2$$

where P and Q represent the propensity score density from benchmark and current survey, respectively.
- Balance measure on covariates
  - Absolute distance

Thank you!

Comments are appreciated!

Contact: julialee@umich.edu