Imputation classes as a framework for inferences from non-random samples. [1]

Vladislav Beresovsky (hvy4@cdc.gov)

National Center for Health Statistics, CDC

[1] Disclaimer: The findings and conclusions in this presentation are those of the author(s) and do not necessarily represent the views of the Centers for Disease Control and Prevention.

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Proliferation of web surveys and the future of NHIS

Is it possible to make reliable estimates from web surveys, planned as extensions to the regular NHIS?

- Current problems with conventional randomized surveys:
  - Increasingly difficult to define frame with good coverage;
  - Growing unit non-response;
  - Expensive data colection.
- Advantages of web surveys:
  - Growing number of vendors offering relatively low cost web panels;
  - Quick turnaround of sampling and data collection;
- Downside of inferences with web survey data:
  - Possibly biased estimates;
  - Unclear how to estimate variances;

CDC
Centers for Disease Control and Prevention

Data description, inferential goals and complications

- Two samples based on NHIS questionnaire collected at the same time (last quarter of 2014):
  - Nonrandom web survey sample s_W. Core questions X_C, detail questions Y_D, unknown weights.
  - NHIS random reference sample s_R. Core questions X_C, no Y_D, known NHIS sampling weights.
- The challenge: How to make inferences for Y_D using X_C as covariates utilizing data from s_W and s_R?
- Similar problem as inferences with missing data and case-control studies. It requires model-based solution.
  - Which model to use: predictive model for Y_D or propensity model of responding to nonrandom web survey?
  - How to account for modeling variability in variance estimation?
  - How to provide for any kind of data analysis with Y_D, rather than just estimating totals?

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Imputation classes framework for inferences from nonrandom samples

Proposed imputation classes framework unifies predictive and propensity models.

- Create imputation classes as areas of homogeneous prediction by **both** propensity and predictive models. Similar idea was advocated for missing data in Haziza & Beaumont *Int. Stat. Rev.* (2007) paper;
- Use hot-deck to impute Y_D from s_W to s_R within these classes;
- Make all kind of estimates from s_R with imputed Y_D and known sampling weights w_r;
- Calculate variances using delete-a-group modification of jackknife with imputed data, described in Rao & Shao *Biometrika* (1992) paper;

CDC
Centers for Disease Control and Prevention

Results of simulation studies

- Very basic simulation study demonstrated:
  - Estimates based on imputation classes are double robust- they are unbiased if just one of the models is correct;
  - Valid inferences for population mean and median using delete-a-group version of Rao & Shao jackknife;
- Simulation study motivated by Kang & Schafer *Statistical Science* (2007) paper "Demystifying Double Robustness ...":
  - When both predictive and propensity models are incorrect, estimates based on imputation classes defined by **both** models are more robust against model misspecification than estimates based on any single model;
  - Estimates by hotdeck within imputation classes are more robust than deterministic imputation;
  - Machine learning have advantage over model-based methods only for large samples or stronger core covariates X_C.

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Awesome power of $\otimes$ response and propensity models

$$\mathrm{BR}\left(\widehat{Y}_{\mathrm{mod}}\right) = \frac{\widehat{Y}_{\mathrm{dir}} - \widehat{Y}_{\mathrm{mod}}}{\widehat{Y}_{\mathrm{dir}} - Y_{\mathrm{pop}}}\ 100\%$$

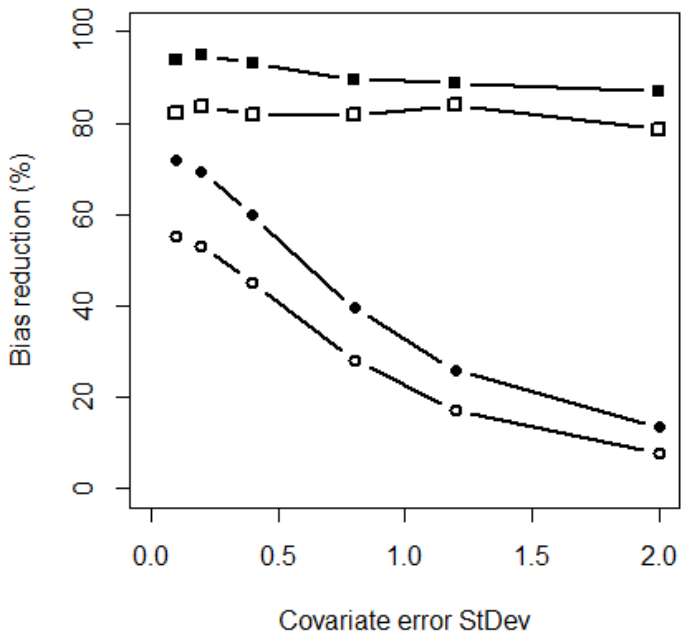

Figure: $\circ - \widehat{\bar{Y}}^{\mathrm{pred}}_{\mathrm{LM.y}}$ $\bullet - \widehat{\bar{Y}}^{\mathrm{PSA}}_{\mathrm{LM.p}}$ $\square - \widehat{\bar{Y}}^{\mathrm{imp5}}_{\mathrm{LM.yp}}$ $\blacksquare - \widehat{\bar{Y}}^{\mathrm{imp10}}_{\mathrm{LM.yp}}$

"It's impossible, probably you are doing something wrong ..."
Julie Gershunskaya, BLS statistician

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

It's all about robustness against model misspecification ...

... Therefore

- Double robust (DR) methods are better than just propensity (PR) or outcome regression (OR) models;
- Robust estimation methods (m-estimation, splines, LASSO, Least Angle Regression, Bayesian spike and slabs, model cross-validation) are preferable at every stage;
- Hotdeck imputation or pair matching techniques are better than deterministic prediction because they may capture information from model residuals (Rubin (1973,1979); Dehejia and Wahba (1999)).
- Biostatisticians have developed great experience working with nonrandom samples in the context of causal inference. There is much to learn from them and to figure out its application to survey samples.

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION